International Journal on Advances in Systems and Measurements

2020 vol. 13 nr. 3&4

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 13, Number 3 & 4, 2020

Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Vítor Carvalho, Minho University & IPCA, Portugal
Irinela Chilibon, National Institute of Research and Development for Optoelectronics, Romania
Soolyeon Cho, North Carolina State University, USA
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noelia Correia, Universidade do Algarve, Portugal
Pierre-Jean Cottinet, INSA de Lyon - LGEF, France
Paulo Estevao Cruvinel, Embrapa Instrumentation Centre - São Carlos, Brazil
Marc Daumas, University of Perpignan, France
Jianguo Ding, University of Luxembourg, Luxembourg
António Dourado, University of Coimbra, Portugal
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France
Matthew Dunlop, Virginia Tech, USA
Mohamed Eltoweissy, Pacific Northwest National Laboratory / Virginia Tech, USA
Paulo Felisberto, LARSyS, University of Algarve, Portugal
Javad Foroughi, University of Wollongong, Australia
Miguel Franklin de Castro, Federal University of Ceará, Brazil
Mounir Gaidi, Centre de Recherches et des Technologies de l'Energie (CRTEn), Tunisie
Eva Gescheidtova, Brno University of Technology, Czech Republic
Tejas R. Gandhi, Virtua Health-Marlton, USA
Teodor Ghetiu, University of York, UK
Franca Giannini, IMATI - Consiglio Nazionale delle Ricerche - Genova, Italy
Gonçalo Gomes, Nokia Siemens Networks, Portugal
Luis Gomes, Universidade Nova Lisboa, Portugal
Antonio Luis Gomes Valente, University of Trás-os-Montes and Alto Douro, Portugal
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Genady Grabarnik,CUNY - New York, USA
Craig Grimes, Nanjing University of Technology, PR China
Stefanos Gritzalis, University of the Aegean, Greece
Richard Gunstone, Bournemouth University, UK
Jianlin Guo, Mitsubishi Electric Research Laboratories, USA
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Petr Hanáček, Brno University of Technology, Czech Republic
Go Hasegawa, Osaka University, Japan
Henning Heuer, Fraunhofer Institut Zerstörungsfreie Prüfverfahren (FhG-IZFP-D), Germany
Paloma R. Horche, Universidad Politécnica de Madrid, Spain
Vincent Huang, Ericsson Research, Sweden
Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek - Hannover, Germany
Travis Humble, Oak Ridge National Laboratory, USA
Florentin Ipate, University of Pitesti, Romania
Imad Jawhar, United Arab Emirates University, UAE
Terje Jensen, Telenor Group Industrial Development, Norway
Liudi Jiang, University of Southampton, UK
Kenneth B. Kent, University of New Brunswick, Canada
Fotis Kerasiotis, University of Patras, Greece
Andrei Khrennikov, Linnaeus University, Sweden
Alexander Klaus, Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
Andrew Kusiak, The University of Iowa, USA
Vladimir Laukhin, Institució Catalana de Recerca i Estudis Avançats (ICREA) / Institut de Ciencia de Materials de Barcelona (ICMAB-CSIC), Spain
Kevin Lee, Murdoch University, Australia
Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway

Andreas Löf, University of Waikato, New Zealand
Jerzy P. Lukaszewicz, Nicholas Copernicus University - Torun, Poland
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Stefano Mariani, Politecnico di Milano, Italy
Paulo Martins Pedro, Chaminade University, USA / Unicamp, Brazil
Don McNickle, University of Canterbury, New Zealand
Mahmoud Meribout, The Petroleum Institute - Abu Dhabi, UAE
Luca Mesin, Politecnico di Torino, Italy
Marco Mevius, HTWG Konstanz, Germany
Marek Miskowicz, AGH University of Science and Technology, Poland
Jean-Henry Morin, University of Geneva, Switzerland
Fabrice Mourlin, Paris 12th University, France
Adrian Muscat, University of Malta, Malta
George Oikonomou, University of Bristol, UK
Arnaldo S. R. Oliveira, Universidade de Aveiro-DETI / Instituto de Telecomunicações, Portugal
Aida Omerovic, SINTEF ICT, Norway
Victor Ovchinnikov, Aalto University, Finland
Telhat Özdoğan, Amasya University - Amasya, Turkey
Gurkan Ozhan, Middle East Technical University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Matteo G A Paris, Universita` degli Studi di Milano,Italy
Vittorio M.N. Passaro, Politecnico di Bari, Italy
Giuseppe Patanè, CNR-IMATI, Italy
Marek Penhaker, VSB- Technical University of Ostrava, Czech Republic
Juho Perälä, Bitfactor Oy, Finland
Florian Pinel, T.J.Watson Research Center, IBM, USA
Ana-Catalina Plesa, German Aerospace Center, Germany
Miodrag Potkonjak, University of California - Los Angeles, USA
Alessandro Pozzebon, University of Siena, Italy
Vladimir Privman, Clarkson University, USA
Mohammed Rajabali Nejad, Universiteit Twente, the Netherlands
Konandur Rajanna, Indian Institute of Science, India
Nageswara Rao, Oak Ridge National Laboratory, USA
Stefan Rass, Universität Klagenfurt, Austria
Candid Reig, University of Valencia, Spain
Teresa Restivo, University of Porto, Portugal
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Harper-Adams University College, UK
Luis Roa Oppliger, Universidad de Concepción, Chile
Ivan Rodero, Rutgers University - Piscataway, USA
Lorenzo Rubio Arjona, Universitat Politècnica de València, Spain
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance, Germany
Subhash Saini, NASA, USA
Mikko Sallinen, University of Oulu, Finland
Christian Schanes, Vienna University of Technology, Austria
Rainer Schönbein, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Germany
Cristina Seceleanu, Mälardalen University, Sweden
Guodong Shao, National Institute of Standards and Technology (NIST), USA
Dongwan Shin, New Mexico Tech, USA
Larisa Shwartz, T.J. Watson Research Center, IBM, USA
Simone Silvestri, University of Rome "La Sapienza", Italy

International Journal on Advances in Systems and Measurements
Volume 13, Numbers 3 & 4, 2020

CONTENTS

Fabien Le Pennec, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France
Amine El Halabi, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France
Sandrine Bernardini, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France
Carine Perrin-Pellegrino, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France
Khalifa Aguir, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France
Marc Bendahan, Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France, France

An Agent-Based Model for Analyzing the HPC Input/Output System

Diego Encinas*†, Sandra Mendez‡, Marcelo Naiouf*, Armando De Giusti*, Dolores Rexachs§and Emilio Luque§
*Informatics Research Institute LIDI. CIC's Associated Research Center.
Universidad Nacional de La Plata. La Plata, 1900, Argentina.
Email: {dencinas, mnaiouf, degiusti}@lidi.info.unlp.edu.ar
†SimHPC-TICAPPS. Universidad Nacional Arturo Jauretche. Florencio Varela, 1888, Argentina.
‡Computer Sciences Department. Barcelona Supercomputing Center (BSC). Barcelona, 08034, Spain.
Email: sandra.mendez@bsc.es
§Computer Architecture and Operating Systems Department. Universitat Autònoma de Barcelona. Bellaterra, 08193, Spain.
Email: dolores.rexachs@uab.es, emilio.luque@uab.es

Abstract—**High Performance Computing (HPC) applications can spend a significant portion of their execution time doing Input/Output (I/O) operations into files. Improving I/O performance becomes more important for the HPC community, as parallel applications produce more data and use more computing resources. One of the methods used to evaluate and understand the I/O performance behavior of such applications in new I/O systems or for different configurations is using modeling and simulation techniques. In this paper, we present a simulation model of the HPC I/O system by using Agent-Based Modeling and Simulation (ABMS) based on the functionality of the I/O Software Stack. Our proposal is modeled using the concept of white box so that the specific behavior of each of the modules or layers in the system can be observed. The interaction between the layers of the I/O software stack are analyzed by monitoring the internal functions using proprietary parallel file system tools. This allows obtaining the functional and temporal characteristics corresponding to the I/O operations. These characteristics allowed the design and implementation of a representative model of I/O system components. Furthermore, measurements are used to obtain the necessary data sets in the verification, fine-tuning and validation stages. The resulting implementation has shown similar behaviors for measured and simulated values when using the IOR benchmark with various file sizes.**

Keywords–Agent-Based Modelling and Simulation (ABMS); HPC-I/O System; Parallel File System.

I. INTRODUCTION

Many scientific applications benefit considerably from the rapid advance of processor architectures used in the modern High Performance Computing (HPC) systems. However, they can spend a significant portion of their execution time doing Input/Output (I/O) operations into files. Inefficient I/O is one of the main bottleneck for scientific applications in a large-scale HPC environment.

In the HPC field, the I/O strategy recommended is the parallel I/O that is a technique used to access data in one or more storage devices simultaneously from different application processes so as to maximize bandwidth and speed up operations. For its implementation, a parallel file system is required; otherwise the file system would probably process the I/O requests it receives sequentially, and no specific advantages in relation to parallel I/O would be gained. Generally, evaluating the performance offered by a HPC I/O system with different configurations and the same application allows selecting the best settings. However, analyzing application performance can also be a useful before configuring the hardware.

One of the methods used to predict the applications behavior under different configurations of the HPC I/O system is using modeling and simulation techniques. That is, analyzing and designing simulation models based on the parallel I/O architecture allows reducing complexity and fulfilling application requirements in HPC by identifying and evaluating the factors that affect performance. In our previous work [1], we presented a methodology for modeling the HPC system, and validated a first simulation design phase focused on components simulation on the client side. Additionally, the code instrumentation method [2] was used to obtain the calibration parameters for the initial version of the simulator. In this work, we expand our model and description by showing the main agents on both client and server sides in a parallel file system. On the other hand, we apply a more accurate method to obtain calibration parameters using system tools to monitor the internal functions of the file system.

In this article, an HPC I/O system is modeled and implemented using the Agent-Based Modelling and Simulation (ABMS) paradigm. The model was built using the I/O software stack functionality. The different layers were ”sensed” by enabling the system's debugging tools. Thus, the necessary data sets were obtained for simulator verification, calibration and validation.

The rest of this paper is organized as follows. Section II briefly describes key I/O concepts, Section III presents the current context of simulation tools for HPC I/O systems, Section IV addresses a functionality analysis for the development of the conceptual model, Section V describes the proposed model, and Section VI describes the computational model of the I/O system. Finally, Section VII presents our conclusion and future work.

II. BACKGROUND

The I/O subsystem in the HPC area consists of two abstraction levels, software and hardware. Usually, the I/O Software includes parallel file system and high level I/O libraries and the I/O hardware refers to servers, storage devices and networks. However, modern HPC I/O system can include more components increasing the complexity of the I/O system.

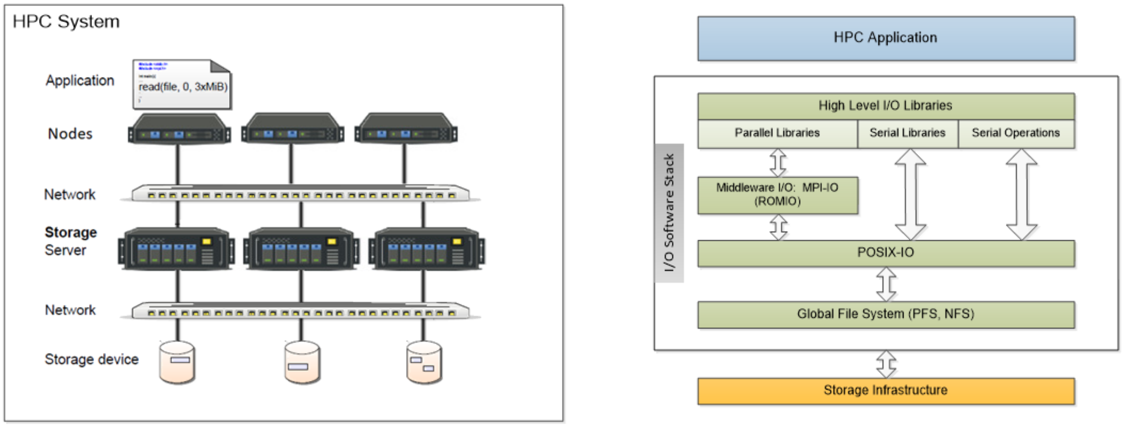


Figure 1. A typical HPC System and the I/O Software Stack

Figure 1 illustrates the structure of the hardware components and the I/O software stack. An I/O operation goes through the software stack from the user application up until it obtains access to the disk from where data are read or on which data are written. Since this parallelism is complex to coordinate and optimize, the implementation of intermediate several layers was designed as a solution.

A. HPC I/O Strategies

The most common I/O strategies in HPC are the serial or parallel accesses into files. Serial I/O is carried out by a single process and it is a non-scalable method because operation time grows linearly with the volume of data and even more with the number of processes, since more time will be required to collect all data in a single process [3].

Parallel I/O usually presents two methods or variations of them: One file per process and a single shared file. In one file per process, each process reads/writes data on its own file on disk and no coordination is required among processes. One single shared file is more convenient to implement Parallel I/O, where all processes write to the same file on disk, but on different sections of that file. This method requires a shared file system that is accessible to all processes.

There are two ways in which multiple processes can access a shared file: independent access and collective access. In the first case, each process accesses the data directly from the file system without communicating or coordinating with the other processes. In collective access, small and fragmented accesses are combined into larger ones to the file system that helps significantly reduce access times. Our aim is to identify this kind of optimizations to explain the I/O behavior, for this reason, we propose a white box model.

B. Middleware

MPI is an interface and communications protocol used to program applications in parallel computers. It is designed to provide basic virtual topology, synchronization, and communication functionalities within a set of processes in an abstract way that is independent from the programming language used to develop the application.

MPI-IO functions work in similar way to those of MPI: writing MPI files is similar to sending MPI messages, and reading MPI files is like receiving MPI messages. MPI-IO also allows reading and writing files in a normal (blocking) mode, as well as asynchronously, to allow performing computation operations while the file on storage device is being read or written on the background. It also supports the concept of collective operations: each process can access MPI files on its own or all together, simultaneously. The second alternative offers greater reading and writing optimizations that can be implemented on several levels. Most of MPI distribution provides MPI-IO functions by using ROMIO [4], which is an implementation of MPI-IO standard and it is used in MPI distributions, such as MPICH, MVAPICH, IBM PE and Intel MPI.

C. Parallel File Systems

A parallel file system is a distributed file system that stripes the files data into multiple data servers, connected to storage devices that provide concurrent access to the files through multiple tasks of a parallel application run on a cluster. The main advantages offered by a parallel file system include a global name space, scalability, and the ability to distribute large files through multiple storage nodes in a cluster environment, which makes a file system like this very appropriate for I/O subsystems in HPC. Typically, a parallel file system includes a metadata server with information about the data found on the data servers.

Some systems use a specific server for metadata, while others distribute the functionality of a metadata server through the data servers. Some examples of parallel file systems for high performance computing clusters are IBM Spectrum Scale, Lustre and PVFS2. PVFS offers three interfaces through which PVFS files are accessed: PVFS' native Application Programming Interface (API), Linux kernel's interface, and ROMIO interface.

The underlying complexity of sending requests to all storage nodes and sorting file contents, among other tasks, is handled by PVFS. When a program attempts a reading operation on a file, small sections of the file are read from several storage devices in parallel. This reduces the load on any given disk controller and allows handling a larger number of requests.

D. Benchmarks

To evaluate the performance of parallel file system and test different I/O libraries of the I/O software stack, there are different I/O benchmarks. Benchmarks are designed to mimic a specific type of workload in a component or system. One the most accepted I/O benchmark in HPC is IOR [5]. It supports several application I/O patterns and allows configuring them, and it offers access to shared files both independently and collectively. Additionally, IOR offers different execution options for the same algorithm using various parallel programming interfaces, including POSIX, MPI-IO, HDF5 and PNetCDF.

III. STATE OF THE ART

There are several research efforts related to HPC I/O system simulators that focus on storage architecture and some layers of the I/O software stack.

The Simulator Framework for Computer Architectures and Storage Networks (SIMCAN) [6] is aimed at optimizing communications and I/O algorithms. The Parallel I/O Simulator of Hierarchical Data (PIOSimHD) [7] was developed to analyze Message Passing Interface-Input/Output (MPI-I/O) performance. The Co-design of Exascale Storage System (CODES) [8] is a framework developed to evaluate the design of the exascale storage systems. The High-Performance Simulator for Hybrid Parallel I/O and Storage System (HPIS3) [9] models application workload. Lustre Simulator [10] was designed to study the scalability of the Lustre file system.

CODES and HPIS3 are based on Rensselaer's Optimistic Simulation System (ROSS) [11], which is a parallel simulation platform. SIMCAN was developed using OMNET++; PIOSimHD was programmed in Java; and Lustre Simulator, in C++. All the tools mentioned use an event-based simulation paradigm (Discrete Event Simulation, DES). We propose using Agent-Based Modeling and Simulation (ABMS) to develop a simulator that will allow evaluating I/O software stack performance.

The agent paradigm is used in various scientific fields and is of special interest in Artificial Intelligence (AI). It allows successfully solving complex problems compared with other classic techniques [12]. It is a simulation technique that recreates the functionality of different components in a real system by modeling entities known as agents. Basically, an agent is an entity capable of perceiving and acting based on changes in its environment. It can also interact with other agents, executing and coordinating its actions, to achieve goals.

Generally, both paradigms operate in discrete time, but DES is used for low to medium abstraction levels. In ABMS, system behavior is defined at an individual level, and global emergent behavior appears when the communication and interaction activities among the agents in an environment start. In fact, ABMS is easier to modify, since model debugging is usually done locally rather than globally [13].

An advantage of ABMS is that different types of models could be created for each part of the system [14][15]. This is useful because the behaviors of the models differ from each other as they are related to diverse actions like processing, communications and storage. Furthermore, environments could be both software- and hardware-based. ABMS allows implementing different components in a modular and flexible way, affording the possibility of connecting and disconnecting different parts of a complex system for a layer-level analysis.

IV. FUNCTIONALITY ANALYSIS

To define an initial model of the I/O system, system functionality should be fully understood. First, the I/O pattern type to be analyzed was selected, and then the corresponding software stack layers for this model were applied. We have selected the IOR benchmark to evaluate I/O performance in HPC clusters. The analysis was focused on the functionality that was observed for IOR in the data path.

Due to the heterogeneity of the I/O systems and the complexity of the software stack, the analysis was started for MPI-IO layers and the parallel file system. PVFS2 was the file system selected for our tests. At this time, we separated the different components considering the concepts of a parallel file system to allow us using the model with other parallel file systems, such as Lustre in the future.

The IOR benchmark offers the total runtime measurements for their programs, but they do not go into further detail in relation to the different abstraction layers of the parallel I/O system. These layers have to be crossed from the moment the user application sends an I/O request up until the CPU, through its operating system, effectively accesses the file on disk to read or write the data. Therefore, it is important to identify the layer in the software stack that requires more time during an I/O operation.

To follow the data path in the software stack, tracers or monitors can be used, but these operate on different levels of the I/O system. There is no single tool that allows recording the I/O behavior in all levels. However, the parallel file systems typically include logging/debugging methods that allows measuring different parameters on the client and server side.

A. Monitoring the internal functions of a parallel file system

The internal functionality of the different components in a HPC I/O system can be identified by: 1) instrumenting the code of the components that are in the data path to perform an I/O operation or 2) using monitoring tools in each level of the software stack. In [1], the code instrumentation in the I/O path was applied to establish what percentage of the total runtime of an I/O operation corresponds to each software stack component. This allows identifying which of them is the most critical one and should be enhanced to improve parallel I/O performance.

The second method requires to monitor the internal behavior of each component of the I/O software stack. As the parallel file system is the I/O software component that is running at client- and server-side, by using its internal logging interfaces, it is possible to identify the internal functionality and its timing for the different component in data path. Some of these tool are Lustre Monitoring Tools (LMT) [16] or Low level Lustre file system configuration utility (lctl) in Lustre [17], or Administration and Monitoring System (AdMon) in BeeGFS [18]. In the case of PVFS2, the options are `gossip` interface and performance counters [19]. In most parallel file systems, these need only to be enabled; they do not require source code re-compilation.

In this paper, we use the PVFS2's `gossip` interface that allows users to specify different levels of logging for the PVFS2 servers. Within the operation principle, `gossip` uses a debugging mask that allows defining which output records are required to print to the log file. Using a global mask, the

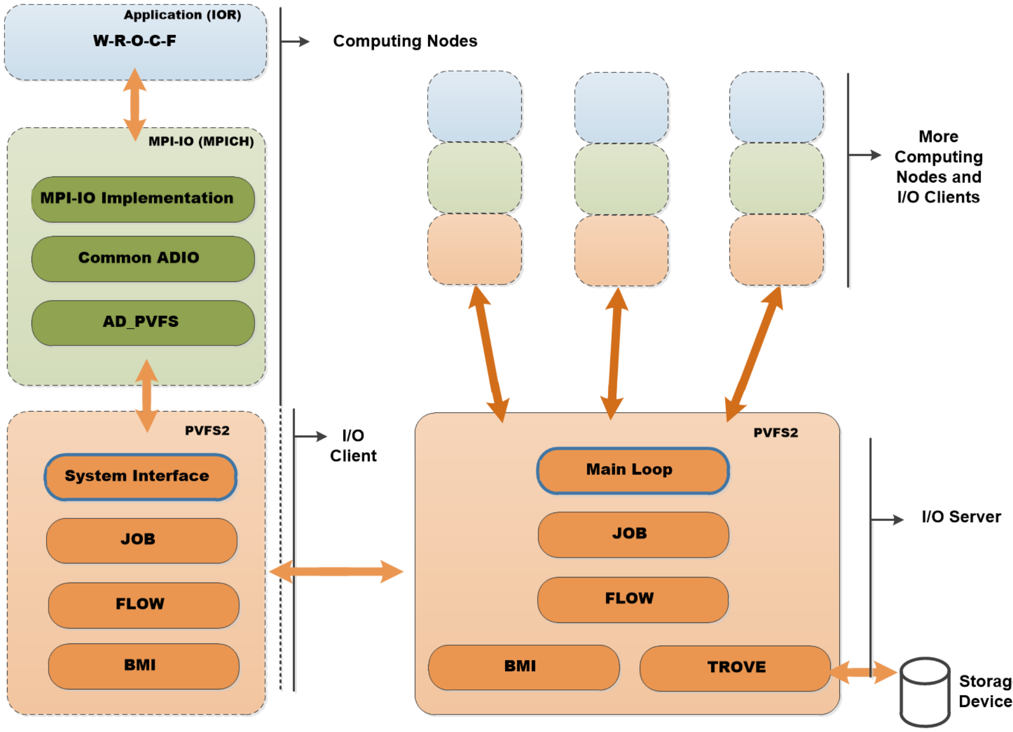


Figure 2. Monitoring in the I/O Software Stack. Left boxes in blue, green and orange represent the layers on compute nodes. The bigger orange box depicts the layers on the I/O Server. Small orange boxes represent the I/O clients, which interact with the metadata and data servers (I/O Servers).

user can specify whether to enable or disable output record sets.

In Figure 2, the different layers of the I/O system can be observed, where the different functions can be measured, both using code instrumentation or the PVFS2's `gossip` interface.

B. Execution Environment

One of the problems found in HPC production systems is that the file system cannot be modified/instrumented, and in most cases, the control of the monitoring level of the internal functionality requires root privileges. Therefore, to deploy scenarios to identify the components functionality, we need to have the total control of the HPC cluster and its I/O subsystem. To create the entire I/O software stack with the appropriate monitoring level, we have deployed a small physical HPC cluster with root privileges and a virtual HPC cluster in the Amazon's EC2 platform.

Platforms like Amazon's EC2 offer various types of instances based on the type of service purchased. In [1], a virtual HPC cluster was deployed using the free service and, even though these nodes offer very limited functionality as regards number of CPUs, memory, storage and network; they proved to be adequate to create the necessary environment for the tests executed. Even though this experiment environment allowed obtaining different measurements to be used in the modeling stage, it has already been mentioned that Amazon's EC2 platform service has restrictions.

Unlike execution environment presented in [1], in this work we present the results obtained in a small physical HPC cluster. However, in both scenarios, it was validated that the observed behaviors follow the same trend even though they do not have the same numerical values.

The deployed I/O configuration has five computing and I/O nodes. In this case, an I/O node fulfills the roles of Client, Data Server and MetaData Server for PVFS2. Through the configuration used, the critical functions involved in each layer of the I/O software stack were selected based on their role and execution time. As way of example, Figure 3 shows the functions selected in the System Interface layer on the client-side and Main Loop layer on the server-side.

V. MODELLING THE I/O SYSTEM

To design a model, the basic characteristics of real system behavior must be obtained first [20]. In this case, the interactions between control, data and communications for basic I/O operations were analyzed: open, read and write. Each operation triggers a succession of interactions that, in turn, initiate different functions such as those shown in Figure 3 in each of the layers of the I/O software stack.

A. System Interactions

The different interactions between client and server to perform read (**r**) and write (**w**) operations are shown in Figure 4. Once both client and server have been initialized, the System Interface layer starts the **r/w** operation. Since every operation that involves communication with Buffered Messaging Interface (BMI), Flow or Trove is considered a Job, a new operation is indicated to the client's Job layer. The Job layer then sends a message to the Flow layer to start a new transmission flow and send a message to BMI, which immediately establishes communication with the server's BMI. In the server's BMI buffer, the message containing the operation to be performed, the job identifier, the associated flow and a BMI client identifier are added.

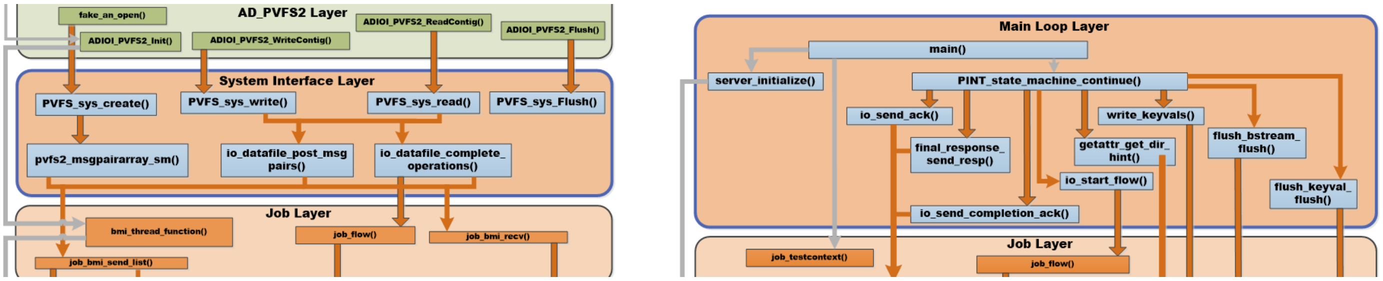


Figure 3. Selected functions of System Interface and Main Loop layers.

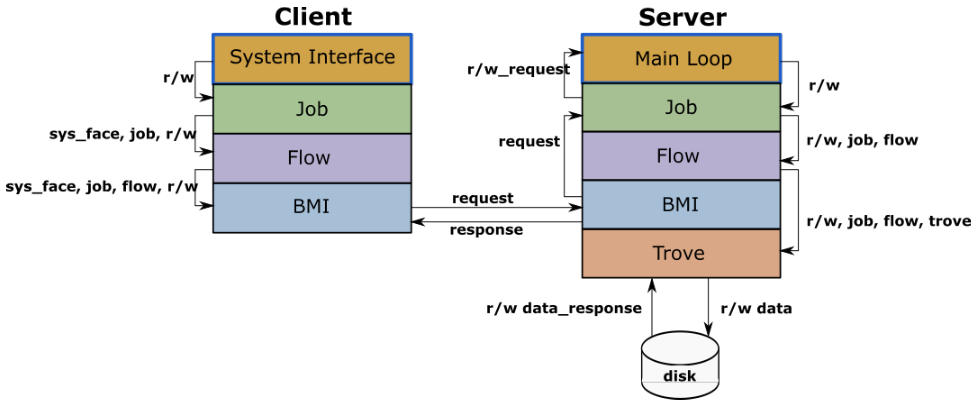


Figure 4. View of the Server-Client Interaction for read (r) and write(w) operations.

On the server-side, once the new operation is detected, the I/O operation is identified and communicated to the Main Loop layer. This layer sends a request to the Job layer to start the new job related to a new flow. Then, a transfer from the Flow layer to disk is carried out. Figure 5 shows how the Flow layer finishes the operation, a response is formally sent to all server layers, and then the server's BMI layer communicates the client's BMI layer that the operation ended.

Figure 4 represents the basic interactions between client and server at the sequential level, but there are other interactions that run in parallel. To carry out an analysis of parallel functionality, the sequence diagram shown in Figure 5 was used. The diagram distinguishes 3 operations: client initializations, server initializations, and the I/O operation itself. As it can be seen, the initializations are run in parallel (highlighted in a blue box for the client and in a green box for the server). Initializations have two purposes: on the one hand, initializing the communications layer on both the client and the server. On the other hand, informing the server that the client is available to establish a communication.

In all interactions, different parameters are sent to each layer in the PVFS2 software stack to identify and carry out the required operation. After the initializations have been performed, the requested I/O operation is executed. The following interactions are sequential and correspond to those mentioned in the description of Figure 4.

B. States Machines

After analyzing each of the layers, a model of the I/O system was developed by implementing state machines and variables that describe each of those states. To that end, state machines were implemented for each of the layers in the system, differentiating their operation both on client- and server-side. The ultimate goal is using these state machines to design the behavior of each of the agents and its interactions with other agents and/or its environment.

The model developed is aimed to reproduce the interaction among the different components and analyzing how the information goes through the different modules or layers, with the possibility of measuring time to approach the real model of the I/O system. Therefore, each layer is modeled based on the execution flow of the functions that are called while processing certain requests, such as opening, closing, reading and writing operations. With the description of each function, the different states of the layers while carrying out those requests were implemented.

Due to the complexity to describe fully the modeling of the I/O software stack, we have selected the System Interface and Main Loop layers to explain in detail the calibration, verification and validation phases. Similar steps were done for the other layers.

The System Interface layer is a client-side interface that allows manipulating the objects in the file system. It launches a number of functions and state machines that process the operation in small steps. In turn, the Main Loop layer is a server layer in charge of controlling whether the operations on lower layers executed by different threads have been completed.

In the context of PVFS, state machines execute a specific function in each of their states. The value returned by this function determines the state that should be adopted. Complex requests can be modeled; they are represented as a sequence of several states. Also, state machines can be nested to model and simplify common subprocess handling. These machines

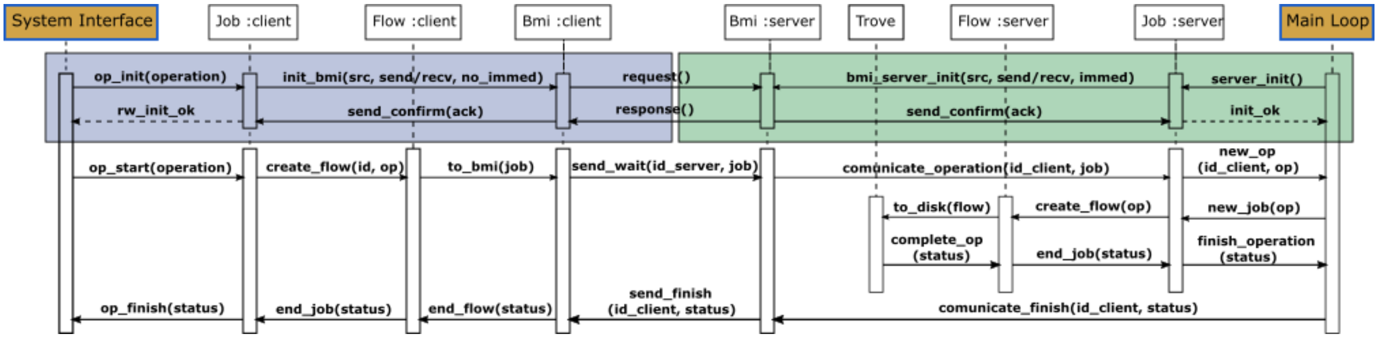


Figure 5. Interaction diagram of System Interface and Main Loop layers for read (r) and write(w) operations.

are used both in clients and servers.

There are several caches on the client side that are part of the System Interface layer and try to minimize the number of requests that the server has to process. The attributes cache (acache) manages metadata, while the name cache (ncache) stores the filename of file system objects and their respective handling number. To prevent caches from storing invalid information, data are set as invalid after a certain time has passed or when the server notifies the client that the object does not exist.

The Main Loop layer accepts four different types of return values related to the invoked operation: completed, deferred, terminated, or failed. It should be noted that the Main Loop layer has one more operation in addition to open, write and read. This is because initialization is an operation in itself, either as a Dataserver or a Metadata server.

C. Functional Model

As shown in Figure 3, the functions in the system interface layer are: `PVFS_sys_create()` to manage the creation of new files in the system, `PVFS_sys_write()` to perform writing operations, `PVFS_sys_read()` to perform reading operations, and `PVFS_sys_flush()` to dump file to data server. The most significant functions of the Main Loop layer include `io_send_ack()`, which returns a negative or positive response to the client; `io_send_completion_ack()`, which reports the completion of an operation that was in progress; and `io_start_flow()`, which initiates a Job to service a Flow depending on the requested operation. Each of these functions has internal variables and state machines that are run to carry out the relevant operations.

Each of these functions has internal variables and state machines that are run to carry out the relevant operations.

To simplify the model, we considered the following in relation to parallelism when handling several instances:

- I/O interfaces: layers `MPI-IO`, `ADIO` and `AD_PVFS` work in a sequential and blocking manner, since they run functions that require synchronization; this means that no instruction is served until the instruction being processed is completed. The calls run on their state machines are blocking;
- PVFS2 parallel file system: the System Interface, Job, Flow, BMI, Main Loop and Trove layers serve other requests and store their instructions in a buffer. Therefore, it allows handling different data flows.

The behavior of each of the agents is described by the state machine, the state transition table and the corresponding state variables. Figure 6 shows part of the state machines developed to model the operation of the System Interface layer, considering the functions and state machines corresponding to each of the three initial operations. As it can be seen, it consists of four agents called System Interface, which is responsible for decoding the instructions that enter the layer; `PVFS_OPEN`, which manages file opening operations; `PVFS_GETATTR`, which carries out searches in the metadata; and `PVFS_RW`, which manages file reading and writing operations.

As way of example, the states of an agent in the System Interface layer in Figure 6 are explained. The agent that manages file opening operations can only have one of five different states (S8 to S12). It will remain in S8 and configure agent `PVFS_GETATTR` if it requests metadata. If the attributes are not found in cache, it will transition to state S9 to wait for them; otherwise, it will transition to state S10. If in state S9, it will wait for a response from agent `PVFS_GETATTR` or it will complete the opening operation by communicating with the server, transitioning to state S10. If the operation cannot be completed, it will transition to state S12 to end.

While in state S10, it will start file creation through a request sent to the JOB layer, transitioning to state S11. Otherwise, it finishes the operation and transitions to state S12. While in state S11, it waits for a response to its file opening request and, if it receives one, it transitions to state S12. Once in state S12, it finishes the operation and sends a response to agent `AD_PVFS`.

Each state of the `PVFS_GETATTR` agent, the same as each of the agents in each layer of the system, has different state variables. These are five per state, and their values depend on their role: ID to identify each process, `DATA_SYSTEM` to indicate permanence in memory or not, `OPERATION` to specify the type of operation, `REQUEST_IN_PROCESS` to indicate if the process has finished or not, and `COD_OPERATION` to add an identifier per traversed layer.

On the other hand, agent `PVFS_RW` manages the write or read requests on client side. In Figure 6, there can be seen in red the functions selected that were used as the basis for the development of each state machine. For example, one of the functions belonging to `pvfs2_msgpairarray_sm()`[21], on which the `PVFS_RW` agent is based, is `io_datafile_post_msgpairs()` that is responsible for managing the data transmissions involved in the creation of files in agent System Interface. These communications occur,

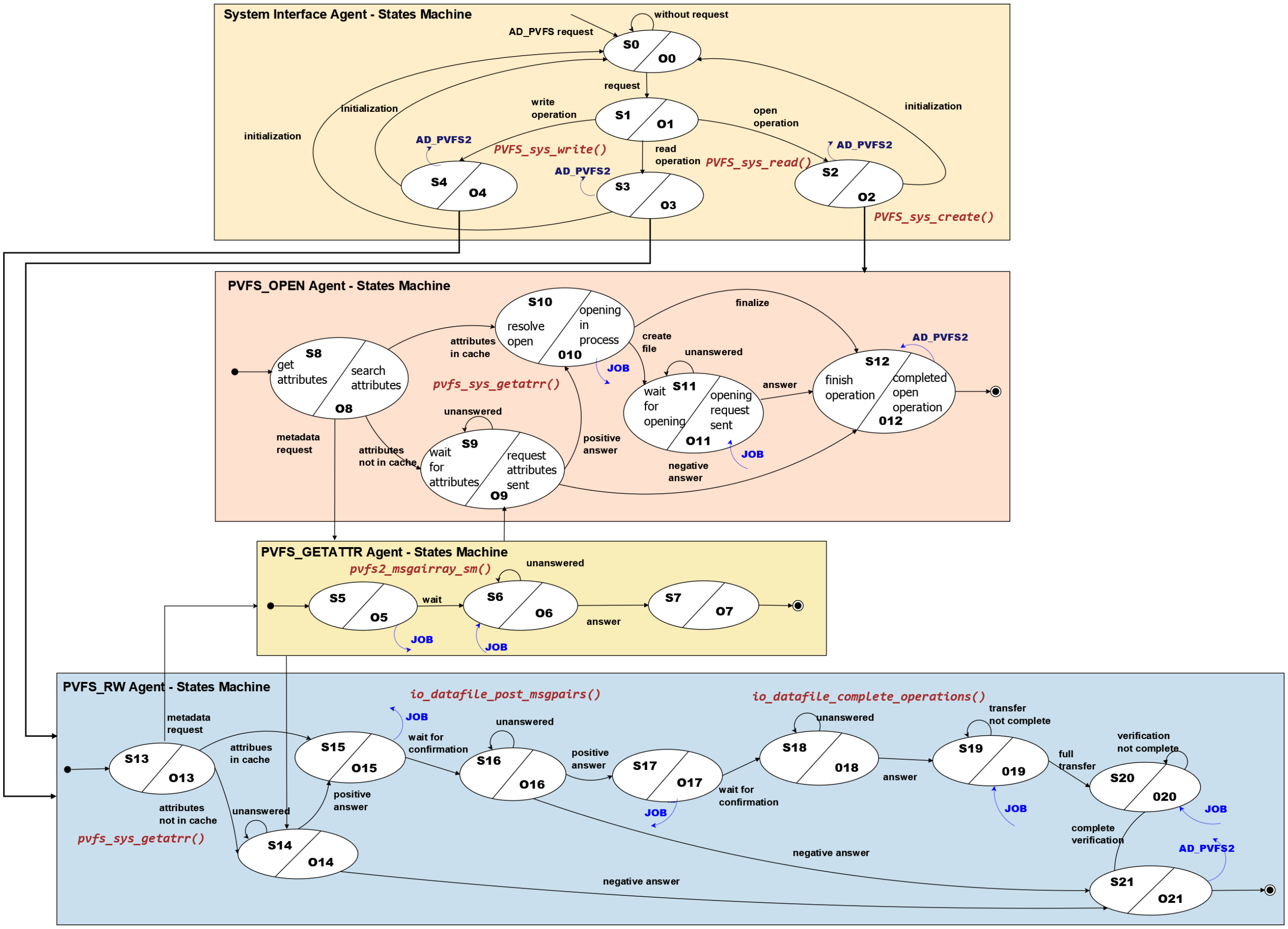


Figure 6. State machines for agents in the system interface layer.

in the case of both a reading or writing, between client and server through the Job and BMI layers.

As for the server's Main Loop layer, Figure 7 shows how it is modeled with 4 agents, namely: MainLoop, which handles server initialization and decodes required operations; MetaData Creation, which reads metadata from disk immediately; File Creation, which writes new files or directories metadata to disk immediately; and Read/Write, which is responsible for configuring data transfers to disk and sending acknowledgment signals to the client.

Figure 7 shows the state machines of each agent, with focus on the states of the Read/Write based on the roles corresponding to pvfs2_io_sm()), which have been marked in red. As previously mentioned, this agent is in charge of managing the data reading or writing operations requested by the client.

VI. COMPUTATIONAL MODEL OF THE I/O SYSTEM

To develop the simulator, tasks were organized in three groups: 1) obtaining data sets that represent the temporary function of the system, 2) using an ABMS-oriented framework, and 3) validating the tool developed.

A. *Verification and Calibration*

To obtain values for the functional model, we have monitored the selected functions for the IOR benchmark in a HPC physical cluster. The I/O system was configured over on PVFS2 parallel file system and the MPICH distribution. The cluster was composed by five nodes, where each one had three roles: compute node (computing and PVFS2 clients), metadata server and data server (datafiles).

We have selected the IOR [5] benchmark as application and it was configured to run a simple pattern for different file sizes and transfer sizes. IOR was configured as follows:

- 1 GiB === mpirun -np 5 ./ior -a MPIIO -b 205m -t 205m -F
- 2 GiB === mpirun -np 5 ./ior -a MPIIO -b 410m -t 410m -F

For this setting, each process writes/reads to/from its own file in transfer sizes defined by the -t parameter. Due to the block size (-b) is equal to the transfer size (-t), only one operation is done by each process. The interface selected was MPI-IO for the *one file per process* (-F) strategy and independent I/O. The mapping corresponds to one MPI process per compute node.

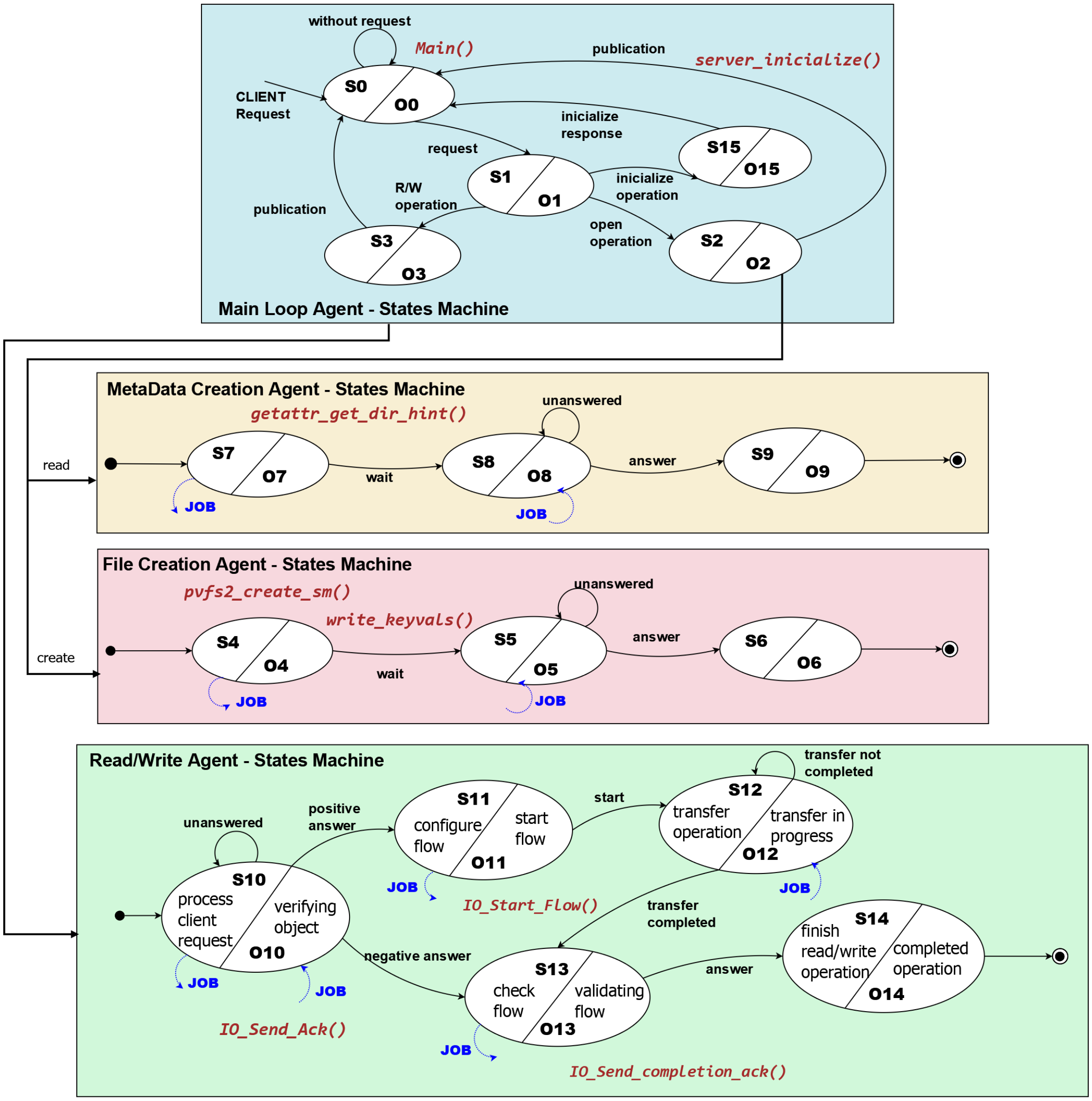


Figure 7. State machines for agents in the main loop layer.

This measurement allows us to classify the monitored metrics in three groups: 1) *data access time* related with the data accesses operations such as write, read, and so on, 2) *control time* that includes verification and configuration of the data structures and 3) *communication time* related with the interaction between the clients and the metadata and data servers.

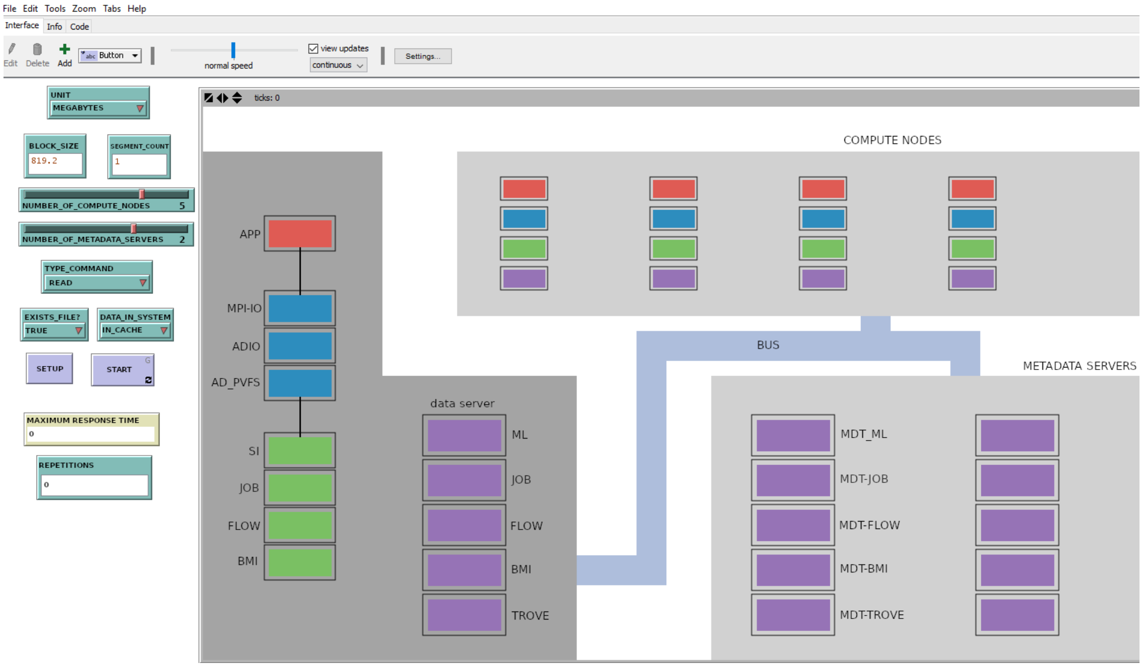
Activating the PVFS2's `gossip` interface the metrics were obtained to apply linear and exponential regressions for the time monitored in different PVFS2's functions. For this analysis, we have selected as dependent variable the execution time and as independent variable the file size, request size is fixed for all the tests. In the case of the system interface layer, we have selected the following equations to represent the time of the functions:

- $\texttt{PVFS\_sys\_create()} = 0.0217x$
- $\texttt{PVFS\_sys\_write()} = 15.183x + 0.0408$
- $\texttt{PVFS\_sys\_read()} = 15.167x + 0.0376$
- $\texttt{io\_datafile\_post\_msgpairs()} = 0.002x^3 - 0.0137x^2 + 0.027x - 3 \cdot 10^{-15}$
- $\texttt{io\_datafile\_complete\_operations()} = -5.6305 \cdot 10^{-7}x^4 + 5.3594 \cdot 10^{-6}x^3 - 1.7401 \cdot 10^{-5}x^2 + 2.1925 \cdot 10^{-5}x + 7.2760 \cdot 10^{-20}$

The equations representing the time functions of the main loop layer are defined as follows:

- $\texttt{io\_start\_flow()}_{read} = 11.3549x$
- $\texttt{io\_start\_flow()}_{write} = 11.4889x$
- $\texttt{io\_send\_ack()} = 3.1987 \cdot 10^{-6}x^3 - 2.4538 \cdot 10^{-5}x^2 + 5.6331 \cdot 10^{-5}x$
- $\texttt{io\_send\_completion\_ack()} = 7.1776 \cdot 10^{-6}x^3 - 5.5622 \cdot 10^{-5}x^2 + 0.00012x$

Where the x variable represents the file size to write or



UNIT: MEGABYTES
BLOCK_SIZE: 819.2
SEGMENT_COUNT: 1
NUMBER_OF_COMPUTE_NODES: 5
NUMBER_OF_METADATA_SERVERS: 2
TYPE_COMMAND: READ
EXISTS_FILE?: TRUE
DATA_IN_SYSTEM: IN_CACHE
MAXIMUM RESPONSE TIME: 0
REPETITIONS: 0

Figure 8. Simulator’s user interface in NetLogo

read. The statistical dispersion also depends on the file size and therefore it is calculated by using the same method.

B. Implementation

The simulation model was developed using an ABMS framework called NetLogo. This framework includes a simplified programming language and a graphical interface that allows the user build, observe and use agent-based modeling without understanding complex standard programming language details. This tool is specifically indicated for the simulation of complex systems; it allows giving instructions to many independent agents that are concurrently executed, which is useful to study the connection between individual and collective behavior through agent actions and interactions.

An implementation detail in this simulator is the use of an agent called ”data” that can be invoked by other agents. This new agent has two main objectives – the first is to calculate the execution time of a function in terms of file size, since the “data” agent can invoke a set of models, algorithms and functions of system components in NetLogo language. The second objective is to generate the simulator output showing the data associated with the invoking agent. Thus, the name of the invoking agent, the associated function based on its state, and the execution time of the function can be displayed.

The scenario adopted for the experiments is similar to the one used in [1], and it was designed to simulate the exchange of information among computing nodes, I/O nodes and storage nodes considering in each of them the layers discussed in previous sections. The MPI operations that can be served by the application layer are only I/O operations, and this initial implementation only includes open, read, write and close operations. One of the parameters allows toggling between executing only one type of operation or all of them. There is an option for selecting a maximum number of operations, which are distributed among the computation nodes selected.

The number of computation nodes and storage nodes can be configured. Node actions and interactions were fully implemented for the operations mentioned above. There are other parameters that allow selecting the existence of the data in the system before running the simulation, configuring the corresponding layers and preparing the I/O server for this scenario.

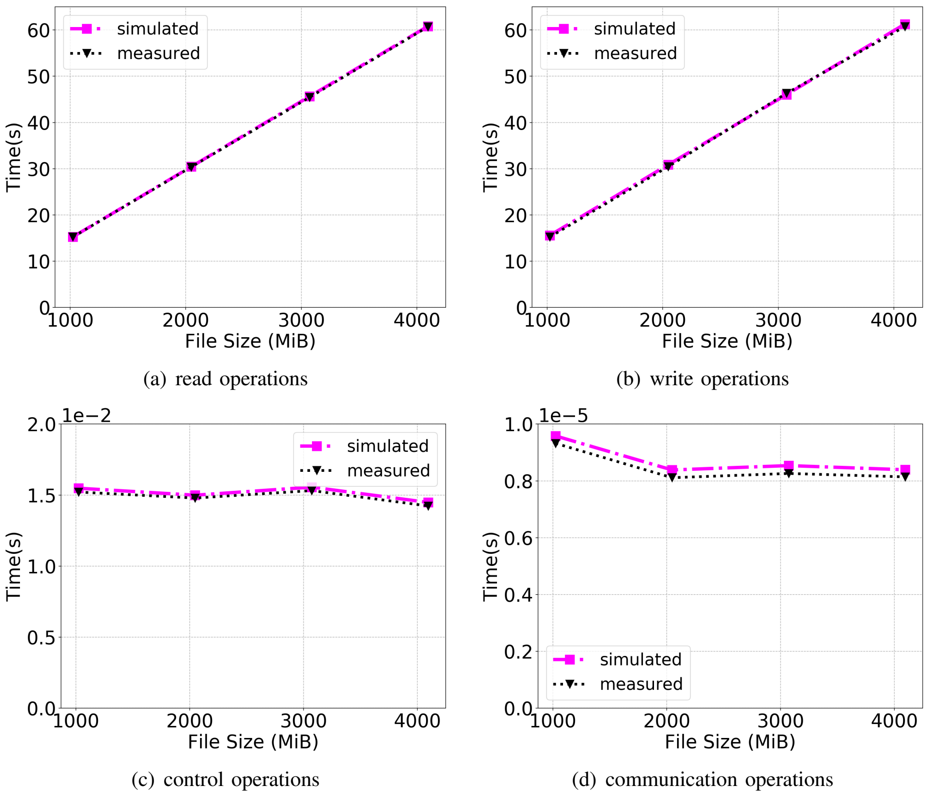
Figure 8 shows the simulator’s user interface. The configuration bars that the user has available to set the variables and parameters of the I/O software stack and the scenario to simulate are on the left. Also, the I/O configuration can be made through command line. The center shows the distribution of the I/O system.

C. Validation

To validate the proposed model, we have configured a physical cluster similar to deployed in the calibration phase (see Section VI-A). The I/O system was deployed by using the PVFS2 parallel file system in a HPC cluster composed by five nodes, where each one was compute node (computing and PVFS2 clients), metadata server and data server (datafiles). PFVS2 filesystem was configured with a stripe size of 64 kiB and a total capacity of 950 GiB. IOR was executed for the following configurations:

- 1 GiB === `mpirun -np 5 ./ior -a MPIIO -b 205m -t 205m -F`
- 2 GiB === `mpirun -np 5 ./ior -a MPIIO -b 410m -t 410m -F`
- 3 GiB === `mpirun -np 5 ./ior -a MPIIO -b 615m -t 615m -F`
- 4 GiB === `mpirun -np 5 ./ior -a MPIIO -b 820m -t 820m -F`

Figure 9 presents the simulated and measured times for the IOR benchmark in the System Interface layer of the PVFS2. As can be seen, the I/O behavior in this layer is dominated



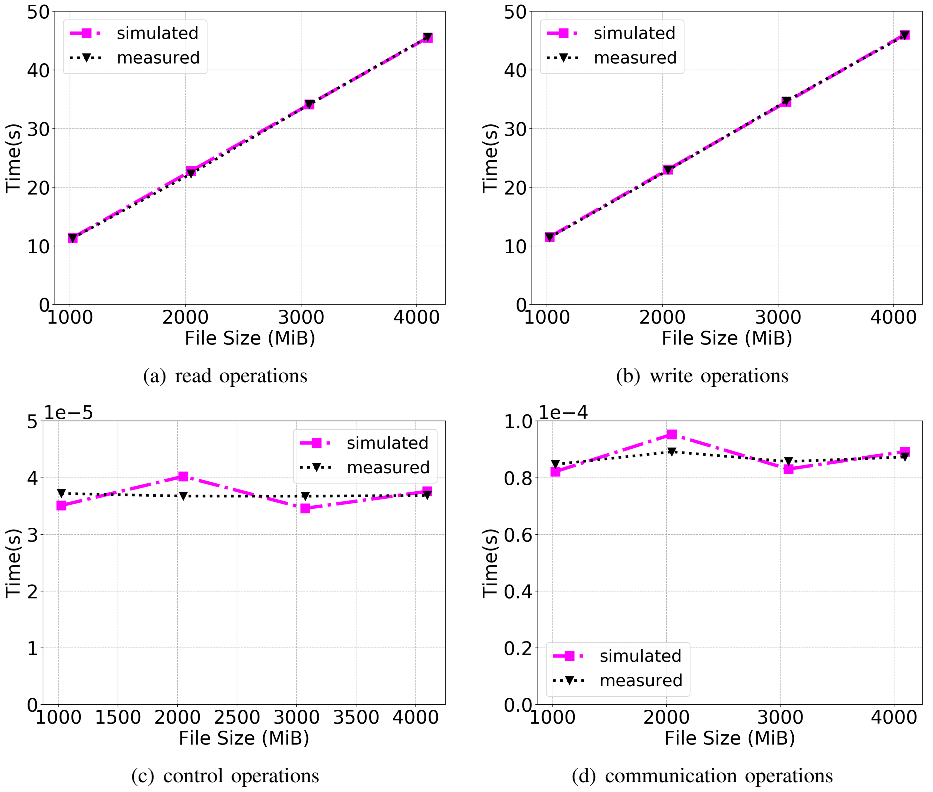
(a) read operations

(b) write operations

(c) control operations

(d) communication operations

Figure 9. Simulated and Measured time for the system interface layer on the PVFS2's client side



(a) read operations

(b) write operations

(c) control operations

(d) communication operations

Figure 10. Simulated and Measured time for the main layer on the PVFS2's server side

by the access data operations that corresponds to the read and write operations. Timings of control and data access operations are very close for 3 GiB and 4 GiB files, which were not tested in the calibration stage. (Section VI-A). Only in the communication operations can be observed a fixed small gap.

Figure 10 shows simulated and measured results at Main Loop level (server-side). Data access operations present a very similar behavior, but we can see different values for the control and communication operations. This is mainly related with functions and constants that are not adjusting perfectly with the real measurements.

The main reason of the accuracy in the measured and simulation results is the simple I/O pattern and the configuration selected. However, this simple HPC I/O system configuration allows us to show that is it possible to model the I/O system behavior properly by using ABMS. Furthermore, from this model, we can deploy different scenarios for the HPC I/O system, including both hardware and software components.

VII. CONCLUSION

This paper presented a model of HPC I/O system by using ABMS, where agents interact and communicate within the I/O software stack layers. To obtain a more representative time for the calibration functions, the interaction between the software stack layers corresponding to the file system were logging with the gossip interface provided by PVFS2. A functional model was defined for the different components of the HPC I/O system by using state machines. The measurement allowed to define equations that represent the temporal behavior for the I/O software stack layers. Furthermore, this was useful for the verification and calibration stages and also for the validation of the simulator developed with the NetLogo modeling environment.

As future work, we will deploy different scenarios for analyzing possible configurations both hardware and I/O software stack. Furthermore, we will evaluate collective operations and other I/O strategies. Additionally, we will extend the model for other parallel file systems, such as Lustre or BeeGFS.

On the other hand, by using the tools for the measurement, we have detected other parameters that can be included in the model and implemented in the simulator, i.e., the data transfer rate (bandwidth) and the input/output operations per second (IOPs).

ACKNOWLEDGMENT

This research has been supported by the Agencia Estatal de Investigación (AEI), Spain and the Fondo Europeo de Desarrollo Regional (FEDER) UE, under contract TIN2017-84875-P and partially funded by the Fundacion Escuelas Universitarias Gimbernat (EUG).

We thank Román Bond, research engineering of Universidad Nacional Arturo Jauretche (Argentina), for his support in the implementation of the simulator.

REFERENCES

[1] D. Encinas, M. Naiouf, A. De Giusti, S. Mendez, D. Rexachs, and E. Luque, "On the Calibration, Verification and Validation of an Agent-Based Model of the HPC Input/Output System," in SIMUL 2019, The Eleventh International Conference on Advances in System Simulation, 2019, pp. 14–21.

[2] D. Encinas, M. Naiouf, A. De Giusti, S. Mendez, D. Rexachs and E. Luque, "Análisis funcional de la pila de software de E/S paralela utilizando IaaS," VI Jornadas de Cloud Computing & Big Data (JCC&BD), 2018, pp. 1–6. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/69465. Retrieved: 08/2020

[3] Sharcnet. Parallel I/O introductory tutorial. [Online]. Available: https://www.sharcnet.ca/help/index.php/Parallel_IO_introductory_tutorial. Retrieved: 10/2019. (2017)

[4] R. Thakur, W. Gropp, and E. Lusk, "Data Sieving and Collective I/O in ROMIO," in Proceedings of the 7th Symposium on the Frontiers of Massively Parallel Computation, ser. FRONTIERS '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 182–189. [Online]. Available: http://dl.acm.org/citation.cfm?id=796733. Retrieved: 10/2019

[5] W. Loewe, T. McLarty, and C. Morrone. IOR Benchmark. [Online]. Available: http://sourceforge.net/projects/ior-sio. Retrieved: 11/2020. (2015)

[6] A. Núñez, J. Fernández, J. D. Garcia, F. Garcia, and J. Carretero, "New techniques for simulating high performance mpi applications on large storage net," J. Supercomput., vol. 51, no. 1, Jan. 2010, pp. 40–57.

[7] J. Kunkel, "Using Simulation to Validate Performance of MPI(-IO) Implementations," in Supercomputing, ser. Lecture Notes in Computer Science, J. M. Kunkel, T. Ludwig, and H. W. Meuer, Eds., no. 7905. Berlin, Heidelberg: Springer, 06 2013, pp. 181–195.

[8] N. Liu et al., "Modeling a leadership-scale storage system." in PPAM (1), ser. Lecture Notes in Computer Science, R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Wasniewski, Eds., vol. 7203. Springer, 2011, pp. 10–19.

[9] B. Feng, N. Liu, S. He, and X.-H. Sun, "HPIS3: Towards a High-performance Simulator for Hybrid Parallel I/O and Storage Systems," in Proceedings of the 9th Parallel Data Storage Workshop, ser. PDSW '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 37–42.

[10] Lustre Simulator. [Online]. Available: https://github.com/yingjinqian/Lustre-Simulator/tree/master/doc. Retrieved: 06/2020. (2016)

[11] C. Carothers, D. Bauer, and S. Pearce, "ROSS: a high-performance, low memory, modular time warp system," in Fourteenth Workshop on Parallel and Distributed Simulation., 2000, pp. 53–60.

[12] V. J. Julián and V. J. Botti, "Estudio de metodos de desarrollo de sistemas multiagente," Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 7, 2003, pp. 65–80. [Online]. Available: http://www.redalyc.org/articulo.oa?id=92501806. Retrieved: 10/2019

[13] A. Borshchev and A. Filippov, "From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools," The 22nd International Conference of the System Dynamics Society, Oxford, England, 07 2004.

[14] E. Kremers, "Modelling and Simulation of Electrical Energy Systems through a Complex Systems Approach using Agent-Based Models," Ph.D. dissertation, Universidad del País Vasco (UPV/EHU), 2012.

[15] M. Taboada, E. Cabrera, F. Epelde, and E. Luque, "Using an agent-based simulation for predicting the effects of patients derivation policies in emergency departments," in International Conference on Computational Science, Barcelona, Spain, 2013, pp. 641–650.

[16] A. Uselton, "Deploying Server-side File System Monitoring at NERSC," in Proceedings of the 2009 Cray User Group, 2009.

[17] Lustre Manual. [Online]. Available: http://doc.lustre.org/lustre_manual.xhtml#idm140436306123424. Retrieved: 06/2020. (2017)

[18] Admon Guide. [Online]. Available: https://www.beegfs.io/wiki/AdmonGuide. Retrieved: 06/2020. (2019)

[19] Fresh Open Source Software Archive. [Online]. Available: https://fossies.org/linux/orangefs/doc/pvfs2-logging.txt. Retrieved: 06/2020. (2017)

[20] A. M. Law, Simulation Modeling & Analysis, 5th ed. New York, NY, USA: McGraw-Hill, 2015.

[21] PVFS2 Team, "PVFS 2 File System Semantics Document," PVFS Development Team, Tech. Rep., 2015.

Design and Objective Evaluation of Filter- and Optimization-based Motion Cueing Strategies for a Hybrid Kinematics Driving Simulator with 5 Degrees of Freedom

Patrick Biemelt, Sandra Gausemeier, and Ansgar Trächtler
Chair of Control Engineering and Mechatronics, Heinz Nixdorf Institute, University of Paderborn, Germany
Email: {patrick.biemelt, sandra.gausemeier, ansgar.traechtler}@hni.uni-paderborn.de

***Abstract*—Dynamic driving simulators have become a key technology to support the development and optimization process of modern vehicle systems both in academic research and in the automotive industry. However, the validity of the results obtained in simulator tests depends significantly on the adequate reproduction of the simulated vehicle movements and the associated immersion of the driver. Therefore, specific motion platform control strategies, so-called *Motion Cueing Algorithms* (MCA), are used to render the acting accelerations and angular velocities within the physical limitations of the driving simulator best possible. In this paper, we describe the design and implementation of two different control approaches for this task, using a simulator with hybrid kinematics motion system as an application example. Motivated by its unique features, an improved filter-based algorithm as well as a real-time capable optimization-based strategy following the idea of *Model Predictive Control* (MPC) are presented and discussed in detail. By means of introduced quality criteria, both algorithms are objectively compared with regard to various standard driving scenarios. These include longitudinal and lateral dynamic maneuvers to estimate the overall improvements of each MCA for interactive driving simulation. Measurement data indicate that both approaches yield an adequate control quality, however, the MPC-based algorithm better handles the kinematic constraints of the simulator due to the integration of additional model knowledge.**

Keywords*–*Interactive Driving Simulation; Motion Cueing; Washout Algorithm; Model Predictive Control; Objective Quality Criteria.

I. Introduction

This article is based on previous work originally presented in [1]. It extends the existing results and provides a deeper understanding of the described concepts and methods.

As a consequence of the constantly increasing multifunctionality and interconnectivity of modern vehicle components and Advanced Driver Assistance Systems (ADAS), automobile manufacturers and developers are facing new technological challenges in recent years. Furthermore, topics such as e-mobility and autonomous driving bring new competitors from the information technology sector onto the market, so that shorter development cycles with simultaneously enhanced product complexities are necessary in order to maintain competitiveness. To overcome those new technological challenges, the use of interactive driving simulators, as shown exemplary in Figure 1, represents an indispensable tool to complement the conventional development process, based on physical prototypes and on-road tests, by model-based test procedures. Such virtual prototyping methods using driving simulators provide the benefit of time and cost savings, as well as safe and reproducible test environments with a high level of flexibility at the same time. For instance, varying weather and lighting conditions can be directly adapted to the test requirements in the simulated environment, which supports i.a. the development and optimization of modern headlamp systems significantly [2]. Furthermore, interactive driving simulation enables to access human-centered aspects, such as marketing, driver training and behavioral studies [3][4].

Figure 1. Interactive Driving Simulators from the Automotive Field [5][6].

Disregarding from the particular analysis purpose, the validity of the results obtained in a virtual test drive is closely linked to the degree of immersion. Interactive driving simulation can therefore be characterized as a *Human- and Hardware-in-the-Loop* (HHiL) application whose transferability to real driving situations can only be guaranteed if a realistic driving impression is created. Hence, it is necessary to provide the human perception system with all required motion information, so-called *Motion Cues*. In addition to the acoustic, visual and haptic stimuli, also the vestibular Motion Cues, more precisely the acting translational accelerations and angular velocities of the simulated vehicle, must be generated using the motion system of the simulator. For this reason, specific Motion Cueing Algorithms are applied in order to create a driving experience that is as realistic as possible within the physical limitations of the motion system.

The most common approach for this task is the *Classical Washout Algorithm* (CWA), which was first described by Schmidt and Conrad as a motion platform control algorithm for piloted flight simulators [7]. As illustrated in Figure 2, this

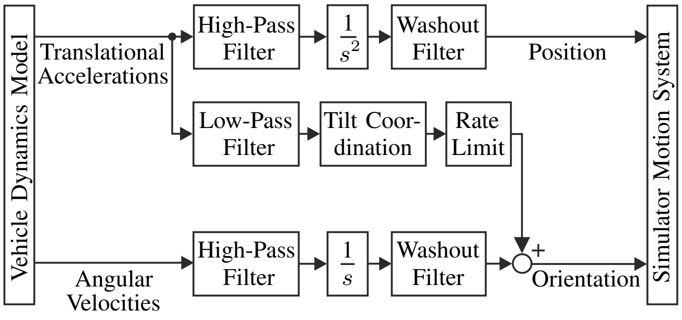


Figure 2. Scheme of the Classical Washout Algorithm [7].

MCA basically consists of a sequence of frequency divisions in order to generate suitable position and orientation reference signals for the simulator motion system. The high-frequency components of the scaled translational accelerations and angular velocities of the vehicle dynamics model are therefore separated using appropriate high-pass filters. Afterwards, these extracted components are directly integrated to a corresponding position and orientation of the driving simulator. Since the basic idea of this algorithm is to return the motion system to its neutral position after it has performed the high-frequency movements, a further high-pass filtering of the integrated signals is conducted. This is known as the *washout effect*. Due to the typically small workspace, an analog integration of the low-frequency accelerations and angular velocities would lead the motion system quickly to its physical limits and thus cannot be performed. Hence, sustained accelerations are simulated via the *tilt coordination* technique, which makes use of the gravitational force to replicate these accelerations by an equivalent rotation of the driving simulator. The corresponding rotation rate is usually limited to the perception threshold of the human vestibular system in this process, so that the rotational motion will not be realized by the driver inside the simulator.

This simple control strategy has been extensively studied and improved since its first publication, typically using hexapod-based motion systems [8][9]. As a result of this research, the filter-based MCA evolved into the standard approach in interactive driving simulation that offers major benefits in terms of transparency and traceability. Each parameter in the Classical Washout Algorithm has a clear physical meaning and a unique association to a single degree of freedom (DOF), which simplifies the tuning significantly. However, this basic idea of treating the translational accelerations and angular velocities independently results in the fact that this approach cannot be applied to every type of motion system. Otherwise, conflicting vestibular stimuli are generated under certain circumstances, e.g., if there exist interdependencies between translational and rotational DOF of the motion system like it is introduced in the next section with the ATMOS driving simulator.

In the present work, we propose an improvement of the CWA that enables a dynamic position washout to any point within the simulator workspace without considerably affecting the high-frequency motion rendering. This key feature is motivated by the considered motion system, but can also be applied to other systems, which offers general advantages for interactive driving simulation. Furthermore, the design and implementation of a real-time capable optimization-based controller is described. It contains additional information by integrating a mathematical model of the motion system, which enables an adequate planning of the simulator trajectory according to the current driving situation. The resulting control quality is evaluated by means of defined objective quality criteria, which take into account both measured and perceived quantities, including models of the human perceptual system. Based on this valuation metric, both MCA are compared using established driving scenarios from the automotive industry, as well as everyday driving maneuvers.

The rest of this paper is structured as follows: Section II provides a detailed overview of the considered motion system and analyzes its specific kinematic characteristics that have to be taken into account to ensure a realistic driving impression. Motivated by these findings, Sections III and IV present the developed filter- and optimization-based MCA. Subsequently, the objective valuation metric and the examined driving scenarios are introduced in Section V, while Sections VI and VII finally discuss the obtained results and give concluding remarks.

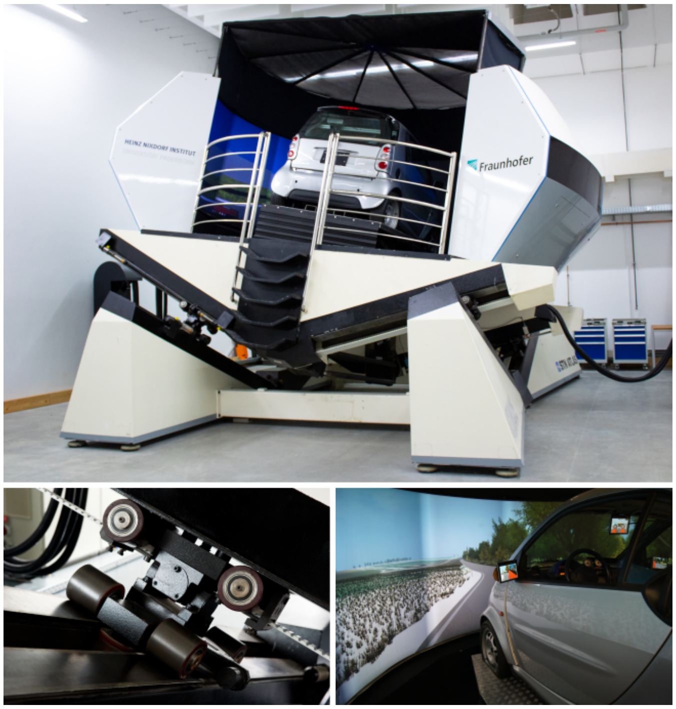


Figure 3. ATMOS Dynamic Driving Simulator.

II. ATMOS Dynamic Driving Simulator

Figure 3 shows the Atlas Motion System (ATMOS) driving simulator that is operated at the Heinz Nixdorf Institute in Paderborn as a reconfigurable development platform, primarily for lighting-based ADAS. As illustrated, this simulator is equipped with a real vehicle chassis of a Smart Fortwo including all its control actuators and instruments, a seamless circular projection with 240 degree viewing angle, a 5.1 multichannel audio system, as well as a unique five DOF motion system to guarantee full immersion of the driver in the virtual environment. Moreover, the acting accelerations and angular velocities are recorded using an Inertial Measurement Unit (IMU) that is installed close to the driver's head position in order to rate the quality of the applied Motion Cueing strategy. In the following, the basic hardware configuration and the dynamic motion system of this simulator will be discussed in detail, as they provide a general understanding of the underlying principles behind the control algorithms presented in Sections III and IV.

A. Simulator Hardware Configuration

To demonstrate its architecture and the interaction of all components within the interactive driving simulation, Figure 4 schematically sketches an overview of the implemented signal and information processing structure. The human driver inside the vehicle chassis, the so-called mockup, forms the core of this simulation setup. With the help of the generated Motion Cues, the driver evaluates the current driving state and performs his steering and pedal inputs to fulfill a specific driving task. Via CAN bus communication, these

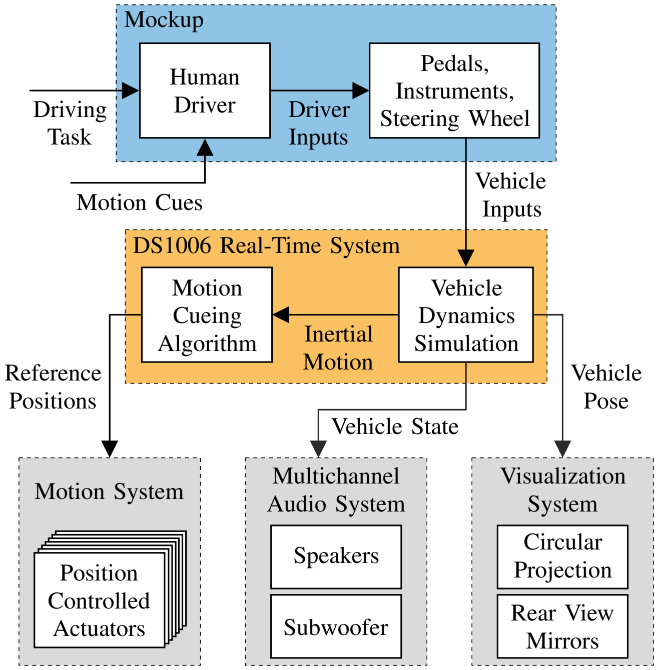


Figure 4. Overview of the Signal and Information Processing.

signals are subsequently processed by a dSPACE DS1006 real-time system using an AMD Opteron CPU @ $2.8\,GHz$, where they serve as inputs for the simulated vehicle in the virtual environment. Here, the *Automotive Simulation Models* (ASM) tool suite is used as vehicle dynamics model, since it is a commercial multibody model that features all relevant subsystems of a real vehicle such as engine, powertrain, axle kinematics, as well as electronic control units and is therefore well-established in automotive applications [10]. The fixed sampling rate thereby is $1\,kHz$, so that all virtual vehicle signals are available without significant latencies. In this way, the computed vehicle pose, consisting of its position and orientation, is determined every millisecond and transmitted to the visualization system. This pose is then displayed with a frequency of $60\,Hz$ on the circular projection, consisting of eight high definition projectors, and three rear view mirror monitors, giving the driver inside the simulator the impression of a fluid movement through the simulated environment. Further information on the applied rendering process of the virtual scenes using the game engine Unity3D is given in [2]. In addition, the characteristic soundscape of the simulated vehicle and other traffic participants is generated according to the calculated vehicle states, such as velocities and engine speeds for example, and reproduced via the installed audio system within the visualization dome. The inertial motion from the vehicle dynamics simulation, specifically the virtual vehicles accelerations and angular velocities, simultaneously serve as an input for the Motion Cueing Algorithm, which is also executed on the real-time system. As described before, the MCA determines suitable control signals for the dynamic motion system to generate the required vestibular stimuli within its physical limitations. In case of the ATMOS driving simulator these control signals contain the reference positions of seven position controlled servo asynchronous motors that drive the system. In the following, the components and the resulting kinematic relations are presented in detail to provide a deeper understanding of this unique motion system.

B. *Dynamic Motion System*

Different from conventional hexapods [11], the motion system of the ATMOS driving simulator is designed as a hybrid kinematics system, which is composed of two mechanically coupled components that can be actuated independently. To illustrate the functionality, Figure 5 shows an exploded view based on the multibody model of the system. The shaker system below the mockup is equipped with three crankshaft drives to perform vertical translational movements, as well as to rotate the driver around the roll and pitch axis. Thus, the shaker replicates the simulated vehicle movements relative to the road surface with exception of yaw motion and can further be used to increase the effect of the tilt coordination by expanding the rotational workspace of the motion system. In addition to the shaker, the motion platform performs movements in lateral and longitudinal direction via four actuated cross-undercarriages that are driven on V-shaped tracks. Because of these tilted tracks, each translational movement of the motion platform leads simultaneously to an additional rotation around the corresponding axes. As a direct consequence of these coupled kinematics, performing pure translational movements of the motion system is only possible within a very small range of the overall workspace, in which the forced rotations of the motion platform can be compensated using the shaker. However, it should be noted that this considerably restricts the shaker systems remaining workspace in its residual degrees of freedom.

To clarify the kinematic properties, the available workspace of the motion platform center point is illustrated in Figure 6. It can be seen that any translational movement causes besides a rotation of the motion platform also a vertical displacement of the center point due to the underlying kinematic constraints.

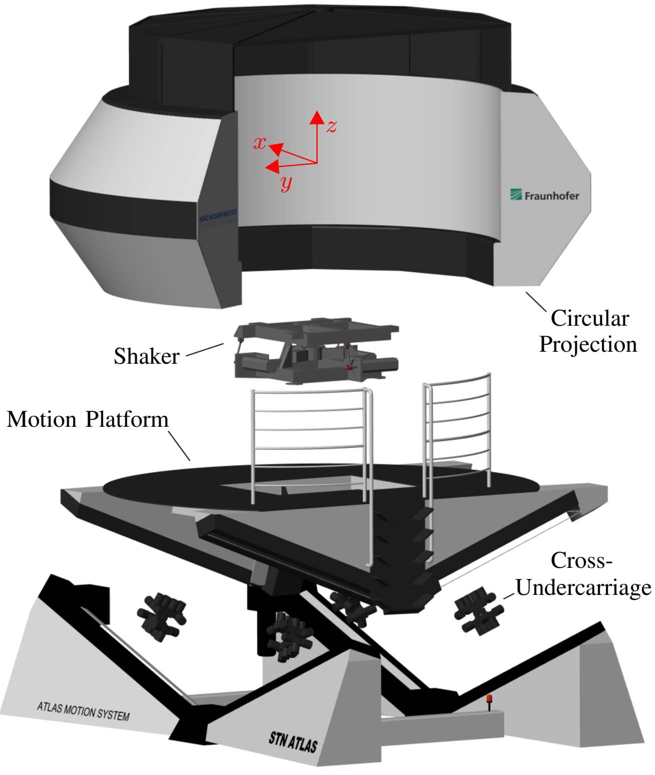


Figure 5. Exploded View of the Simulator Multibody Model.

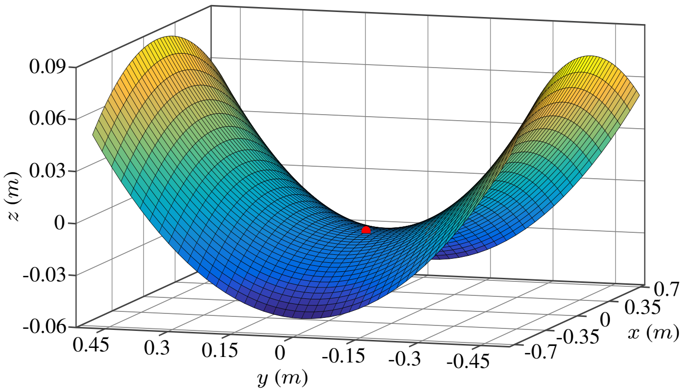


Figure 6. Workspace of the Motion Platform Center Point.

Thus, longitudinal movements always cause a lowering of the platform center, while lateral movements lift it. As a consequence, the motion platform performs movements along the curved surface shown in Figure 6, leading to an additional kinematic coupling between the translational DOF. Analogously, the analysis of the available shaker workspace leads to the dependencies between vertical displacements z, roll inclinations φ and pitch inclinations θ presented in Figure 7. As shown, a maximum vertical displacement of $z = \pm 72\,mm$ is feasible with the shaker. However, this is only practicable if there are no simultaneous tilts of the system, since additional roll and pitch angles not equal to zero considerably reduce the vertical workspace. Roll movements are generated by an alternating actuation of both front crankshaft drives, which are installed symmetrically to the roll axis. Thus, also a symmetric workspace results, as it is pictured top left in Figure 7. In contrast, pitch rotations are generated by actuating the two crankshaft drives in the front and the crankshaft drive in the rear in opposite directions. Due to the geometric properties of the system, the rear actuator reaches its top or bottom dead center at an angle of $\theta = \pm 5°$. A tilt up to the maximum pitch angle of $\theta = \pm 7°$ is then possible by further movements of the front two actuators, but this simultaneously leads to a lifting or lowering of the shaker platform, as shown in the upper right corner of Figure 7. As a consequence, an asymmetrical workspace results. The combination of both upper graphics leads to the overall workspace of the shaker illustrated in the bottom of Figure 7. It shows that there are also interdependencies between the individual DOF of the shaker system, which can in the case of pitch rotations even cause undesired vertical movements of the driver in the simulator. Together with the nonlinear kinematic properties of the motion platform, these aspects has to be considered in the design of the Motion Cueing Algorithm in order to avoid conflicting sensory information, so-called *False Cues*, which typically lead to the undesired effect of *Simulator Sickness* for the driver [12].

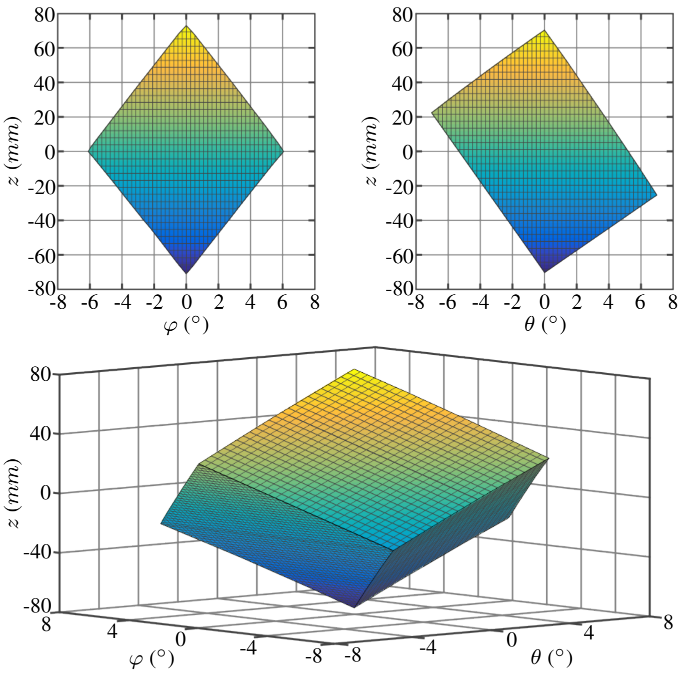


Figure 7. Analysis of the Shaker System Workspace.

Thus, due to the mentioned features of the ATMOS driving simulator, suitable control strategies are required since the implementation of the conventional CWA according to Figure 2 does not result in the desired quality of the motion rendering.

III. Modified Washout Algorithm

As described in Section I, the general idea of the Classical Washout Algorithm is based on an independent consideration of the systems degrees of freedom, which is due to the fact that the MCA was developed for application on a conventional hexapod. Because of this, the algorithm is not suitable for application on the ATMOS driving simulator introduced in the previous section, as there is a connection between translation and rotation because of the underlying kinematics of the motion system. For this reason, we subsequently present an extension of the classical approach that includes the relevant kinematic effects and enables a sufficient control quality. Moreover, a further analysis using system theoretical methods is described in [13].

A. Dynamic Position Washout

In case of the regarded driving simulator, each longitudinal and lateral movement of the motion platform generates a forced tilting around the corresponding roll and pitch axis. These rotations should ideally be used to emulate sustained accelerations using the tilt coordination technique. Otherwise, the tilt coordination has to be performed only by the shaker, which limits the maximum possible inclination to the small shaker workspace (see Figure 7). In contrast to the classical algorithm, a dynamic position washout is therefore required that enables the motion platform to drift into a defined end position within its workspace after it has performed the high-frequency movements. By determining this end position according to the associated inclination, low-frequency accelerations can also be simulated via the motion platform. For this purpose, the high-pass (hp) and washout (wo) filters of the high-frequency longitudinal and lateral acceleration paths are supplemented by further first order low-pass filters with variable gains K, as shown in Figure 8 using the example of longitudinal acceleration a_x. According to the shown structure, the corresponding transfer function G, that describes the dynamic behavior between the acceleration input a_x and the longitudinal simulator position x, is given as

$$G(s) = \frac{T_{hp}s + K}{T_{hp}s + 1} \cdot \frac{T_{wo}^2 s^2}{T_{wo}^2 s^2 + 2DT_{wo}s + 1} \cdot \frac{1}{s^2} \ . \qquad (1)$$

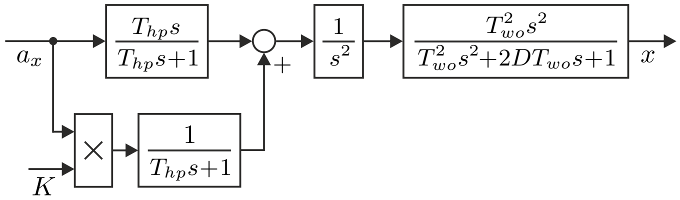

a_x
$\frac{T_{hp}s}{T_{hp}s+1}$
$+$
$\frac{1}{s^2}$
$\frac{T_{wo}^2 s^2}{T_{wo}^2 s^2+2DT_{wo}s+1}$
x
$\times$
$\frac{1}{T_{hp}s+1}$
K

Figure 8. Extended Longitudinal High-Frequency Acceleration Path.

The non-intuitive idea of this extension can be clarified by the application of the final value theorem of the Laplace transform. Therefore, let a_x be a sustained acceleration input from the vehicle dynamics simulation, which can be assumed to be approximately constant, since the magnitude does not significantly change. For the integrated simulator position x follows then with increasing time $t \to \infty$:

$$\lim_{t\to\infty} x(t) = \lim_{s\to 0} s \cdot \underbrace{G(s) \cdot \frac{a_x}{s}}_{X(s)} = K \cdot T_{wo}^2 \cdot a_x \tag{2}$$

Consequently, the resulting simulator position depends on the gain K, the time constant T_{wo} of the washout filter as well as the amplitude of the acting acceleration a_x. If this position is now required to have a defined value x_{tc}, the necessary gain K can be determined corresponding to (2) as

$$K = \frac{x_{tc}}{T_{wo}^2 \cdot a_x} \, . \tag{3}$$

Here, the singularity occurring for $a_x = 0\, m/s^2$ is not critical, since in this case the entire transfer function G is also multiplied with this input variable, resulting in a position $x = 0\, m$. The overall stability of the proposed structure is therefore always guaranteed as long as high-pass and washout filters possess a stable pole configuration, which is generally to be expected. Analogously, the initial value theorem of the Laplace transform can be used to show that the extension by the variable gain low-pass filter, as shown in Figure 8, does not negatively affect the reproduction of high-frequency acceleration components [13]. Like in the Classical Washout Algorithm, the dynamics of the drift into the end position x_{tc} can be specified by the parameters of the washout filter, which represents an important design freedom in the parameterization of the proposed control strategy.

The described extension is also implemented for the lateral high-frequency acceleration path, so that a washout in the defined position y_{tc} analogue to (3) is realized and thus sustained lateral stimuli are produced by a corresponding roll rotation of the motion platform.

B. Tilt Coordination Distribution

Due to the hybrid kinematics motion system, as well as the presented dynamic position washout, the tilt coordination technique can be performed either using the motion platform (mp), the shaker (sh) or a combination of both systems. The latter significantly increases the workspace and thus the maximum low-frequency acceleration amplitudes that can be generated. Consequently, a distribution strategy has to be specified, which enables a suitable coordination of both components. For this reason, an adaptation of the low-frequency longitudinal and lateral acceleration paths is conducted according to Figure 9. As shown with the example of the longitudinal acceleration, a first order low-pass (lp) filter extracts the sustained acceleration components from the reference signal a_x, which are subsequently converted to the corresponding tilt coordination pitch angle θ_{tc}. In doing so, the associated rotation rate is limited to the well-established value of $0.1\, rad/s$, in order that the tilt coordination technique does not disturb the driving impression of the human driver [14]. In contrast to conventional hexapods, this inclination is divided among the subsystems of the motion system by introducing a distribution coefficient $\alpha \in \mathbb{R}$ with $0 \leq \alpha \leq 1$. This results in the inclinations for the shaker θ_{sh} and for the motion platform θ_{mp} that are necessary to replicate the low-frequency accelerations by the gravitational force. Based on the known kinematic relations of the motion platform, an equivalent platform position x_{tc}, which corresponds to the required inclination, is subsequently determined. This position equivalent then serves as input for calculating the variable gain K according to (3) so that the coupling between translational and rotational DOF is taken into account. Equally, this process is implemented for the lateral low-frequency acceleration path.

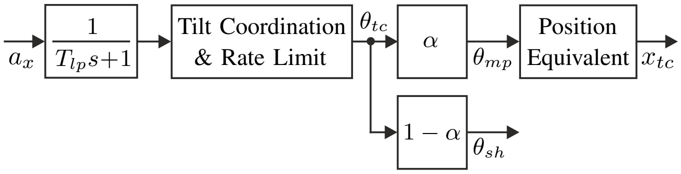


Figure 9. Extended Longitudinal Low-Frequency Acceleration Path.

C. Resulting Algorithm Structure and Parameterization

The combination of dynamic position washout and tilt coordination distribution leads to the overall structure of the modified washout algorithm illustrated in Figure 10. Based on the principles of the Classical Washout Algorithm, this filter-based control strategy enables the generation of suitable control signals in the form of position and orientation commands for the motion system of the ATMOS driving simulator. Using the inverse kinematics of the motion platform and the shaker, the required reference angles of the position controlled actuators are determined, enabling the motion system to generate the vestibular Motion Cues according to the current driving situation. In order to ensure that these references are adjusted to the system with a desired dynamic behavior, a model-based approach to compensate existing actuator latencies is presented in [13]. The estimation of the associated filter parameters and distribution coefficients was performed by numerical optimization using a defined driving maneuver. Here, the rural road drive, which will be introduced in one of the next sections, was chosen since it represents a good compromise between moderate driving scenarios and extreme maneuvers at the limits of driving dynamics. Table I provides an overview of the resulting parameters.

Although the developed algorithm is motivated by the specific features of the motion system, in particular the concept of a dynamic position washout offers great potential to transfer and combine it with alternative Motion Cueing approaches. For example, an integration of the approach into predictive algorithms is possible in order to use information of the current vehicle state and the oncoming road conditions to preposition the motion system. Thus, the available simulator

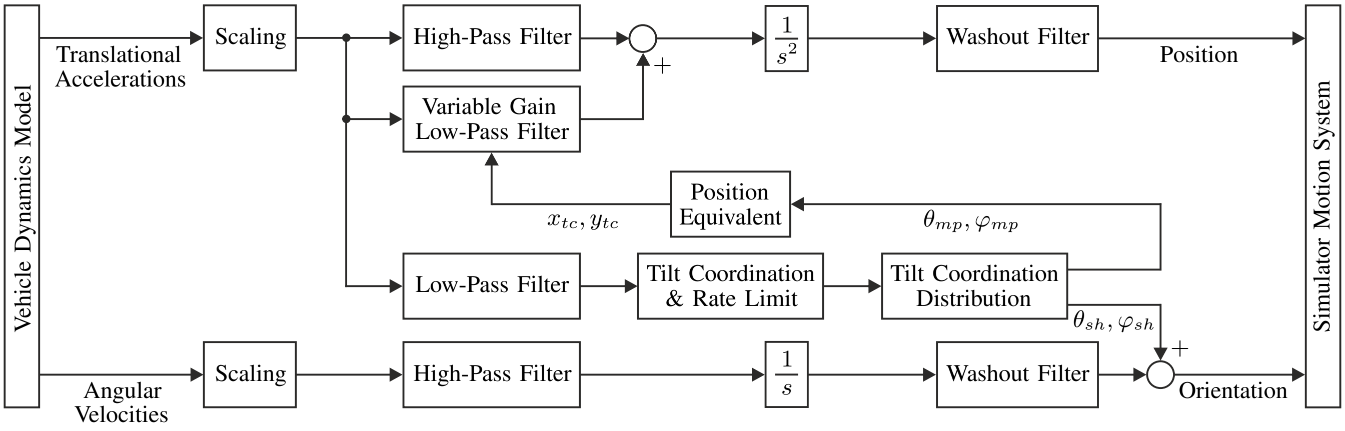


Figure 10. Overall Structure of the Developed Washout Algorithm.

TABLE I. APPLIED ALGORITHM PARAMETERS.

Scaling	*1st Order HP Filter*	*1st Order LP Filter*	*2nd Order WO Filter*	*Distribution Coefficient*
$k_x = 0.4$	$T_{hp} = 0.95$	$T_{lp} = 0.95$	$T_{wo} = 0.49$, $D = 0.7$	$\alpha_x = 0.65$
$k_y = 0.4$	$T_{hp} = 0.6$	$T_{lp} = 0.6$	$T_{wo} = 0.44$, $D = 1.0$	$\alpha_y = 0.6$
$k_z = 1.0$	$T_{hp} = 0.4$	–	$T_{wo} = 0.45$, $D = 1.0$	–
Scaling	*1st Order HP Filter*	*1st Order LP Filter*	*1st Order WO Filter*	*Distribution Coefficient*
$k_\varphi = 1.0$	$T_{hp} = 1.2$	–	$T_{wo} = 0.8$	–
$k_\theta = 1.0$	$T_{hp} = 0.3$	–	$T_{wo} = 0.2$	–

workspace is used more efficiently [15]. For this purpose, suitable positions are determined at runtime instead of x_{tc} and y_{tc}, to which the motion system drifts after executing the high-frequency movements. Occurring false cues caused by the dynamic position washout can thereby be masked by the gravitational force using an additional tilt of the driving simulator [9].

IV. MODEL PREDICTIVE CONTROL APPROACH

While the presented modified washout algorithm takes into account coupling effects between translational and rotational DOF of the ATMOS driving simulator, this filter-based control strategy does not consider interdependencies between the particular translational movements. That can be explained by the underlying algorithm structure, which is basically comparable to the CWA with its independent treatment of all system degrees of freedom. To overcome this, an optimization-based Motion Cueing Algorithm using the concept of Model Predictive Control was introduced in [16]. It offers the advantage that hard constraints, such as the workspace limitations and kinematic relations described in Section II, can be explicitly integrated into a numerical optimization process, which is performed at runtime. Furthermore, by including an actuator dynamics model it is ensured that the determined motion trajectory is always feasible for the driving simulator. In the following, the main aspects of the MPC-based algorithm are explained in detail to provide a basic understanding for the comparison of both control approaches in the next section.

A. Nonlinear Motion System Model

According to the basic idea of the established MPC paradigm, a constrained optimal control problem is numerically solved over a receding time horizon at each calculation cycle. Subsequently, only the first element of the computed trajectory is applied to the process and the procedure is iterated [17]. Thereby, the resulting control quality depends significantly on the availability of an adequate process model to predict the future system behavior. This model consequently has to cover all relevant dynamic and kinematic effects on the one hand. At the same time an online optimization causes a significant computational effort, for which reason the integrated system model must be designed as simple as possible to meet the real-time requirements.

Driving simulators are large-scale systems with high inertia, so there is always a specific dynamic behavior, which influences the control quality and therefore has to be considered in the planning of the simulator motion trajectory. Assuming that the basic mechanical system is a rigid body without significant elasticities, the overall system dynamics can be expressed by the transfer behavior of the installed actuators. In case of the considered motion system, the input/output dynamics of each position controlled actuator is described by a linear third order lag element with the state space representation

$$\begin{aligned} \dot{x}_s(t) &= A_s \cdot x_s(t) + B_s \cdot u_s(t) \\ y_s(t) &= C_s \cdot x_s(t)\ . \end{aligned} \tag{4}$$

Here, the associated state vector $x_s(t) \in \mathbb{R}^3$ contains the angle $\psi(t)$ of a servo motor, its angular velocity $\dot{\psi}(t)$ and its angular acceleration $\ddot{\psi}(t)$:

$$x_s(t) = \begin{bmatrix} \psi(t) & \dot{\psi}(t) & \ddot{\psi}(t) \end{bmatrix}^T \tag{5}$$

The input and output variables of the model from (4) form the reference position $\psi_{ref}(t)$ determined from the MCA and the actual angle $\psi(t)$ of the controlled actuator:

$$\begin{aligned} u_s(t) &= \psi_{ref}(t) \\ y_s(t) &= \psi(t) \end{aligned} \tag{6}$$

Consequently, the state differential equation matrices result as

$$A_s = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix} \in \mathbb{R}^{3\times3},\quad B_s = \begin{bmatrix} 0 \\ 0 \\ b_0 \end{bmatrix} \in \mathbb{R}^3, \tag{7}$$

so that a time-invariant *Single-Input-Single-Output* (SISO) system in controllable canonical form is given for each controlled actuator. As already explained in Section II, the motion system of the ATMOS driving simulator is equipped with a total number of seven servo motors. Three of these are used to actuate the shaker, while two identical servo motors, which are controlled with the same reference positions $\psi_{ref}(t)$, each drive the motion platform in longitudinal and lateral direction. Therefore, in the derived simulator model, the four actuators of the motion platform can be combined to one actuator per longitudinal and lateral DOF to reduce the resulting model complexity. Summarizing all five actuator models finally leads to a 15th order linear system with the state differential equation

$$\dot{x}(t) = A \cdot x(t) + B \cdot u(t)\,. \tag{8}$$

Since the underlying position controllers are very exact and the actuators can thus be assumed to be completely decoupled, the state matrix $A \in \mathbb{R}^{15\times 15}$ and the input matrix $B \in \mathbb{R}^{15\times 5}$ are block diagonal matrices that contain the state differential equations according to (4) of all five servo motors on their main diagonals. The corresponding state vector $x(t) \in \mathbb{R}^{15}$ results from the state variables of each actuator given in (5), while the input vector $u(t) \in \mathbb{R}^{5}$ is a vector obtained from the respective reference positions $\psi_{ref}(t)$.

In order to respect the relevant kinematic characteristics of the simulator explicitly in the control algorithm, a functional relationship between the state variables of (8) and the control variables, more precisely the acting translational accelerations $a(t)$ and angular velocities $\omega(t)$, is required. Moreover, these output quantities need to be described at the driver's head position since the vestibular perception organs are located in the human inner ear [18].

For this purpose, the direct kinematics of the motion system are defined in Cartesian coordinates as

$$\begin{aligned} {}_I r_h\left(\psi(t)\right) &= {}_I r_{mp}\left(\psi(t)\right) + {}_I r_{sh}\left(\psi(t)\right) \\ {}_I \beta_h\left(\psi(t)\right) &= {}_I \beta_{mp}\left(\psi(t)\right) + {}_I \beta_{sh}\left(\psi(t)\right) \end{aligned} \tag{9}$$

in the first instance. According to Figure 11, the pose of the driver's head position h is given by the position and orientation vectors ${}_I r_h = {}_I\left[x\ y\ z\right]^T \in \mathbb{R}^3$ and ${}_I\beta_h = {}_I\left[\varphi\ \theta\right]^T \in \mathbb{R}^2$ in the inertial reference frame I. These are expressed as functions of all five actuator angles $\psi(t)$, which form the systems generalized coordinates in that context. Because the mechanical coupling between the motion platform and the shaker represents a serial kinematics, the positions and orientations of both subsystems are added as shown in (9). To obtain the associated translational and angular velocities, the time derivatives of both vectors are determined:

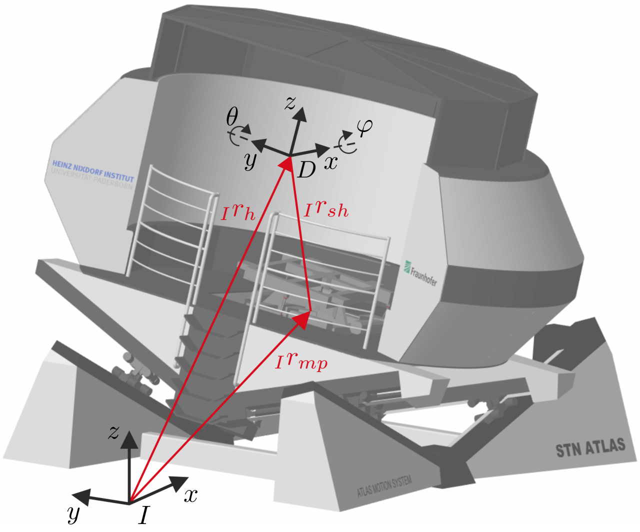


Figure 11. Scheme of the Driver's Head Position Pose.

$$\begin{aligned} {}_I v_h(\psi(t),\dot{\psi}(t)) &= \frac{d_I r_h\left(\psi(t)\right)}{dt} = \frac{\partial_I r_h\left(\psi(t)\right)}{\partial\,\psi(t)} \cdot \dot{\psi}(t) \\ {}_I \dot{\beta}_h(\psi(t),\dot{\psi}(t)) &= \frac{d_I \beta_h\left(\psi(t)\right)}{dt} = \frac{\partial_I \beta_h\left(\psi(t)\right)}{\partial\,\psi(t)} \cdot \dot{\psi}(t) \end{aligned} \tag{10}$$

Hence, the velocity variables of the driver's head position are calculated from the product of the actuator angular velocities $\dot{\psi}(t)$ and the partial derivatives of (9) to the generalized coordinates $\psi(t)$, which is known as the *Jacobian matrix*. A further differentiation of the velocity vector ${}_I v_h(t)$ then yields the desired expression of the translational accelerations ${}_I a_h = {}_I\left[\ddot{x}\ \ddot{y}\ \ddot{z}\right]^T$ according to

$$\begin{aligned} {}_I a_h(\psi(t),\dot{\psi}(t),\ddot{\psi}(t)) &= \frac{d_I v_h(\psi(t),\dot{\psi}(t))}{dt} \\ &= \frac{\partial^2{}_I r_h(\psi(t))}{\partial\,\psi(t)^2} \cdot \dot{\psi}^2(t) \;+\; \frac{\partial_I r_h(\psi(t))}{\partial\,\psi(t)} \cdot \ddot{\psi}(t)\,. \end{aligned} \tag{11}$$

As shown, besides the state variables of (8) and the Jacobian matrix, also the second partial derivatives of the position vector ${}_I r_h(t)$ to the actuator angles $\psi(t)$ are required to determine the acting accelerations at the driver's head position. In addition, the angular velocity vector ${}_I\omega_h(t)$ is obtained from the derivatives of the orientations ${}_I\dot{\beta}_h(t)$ according to (10) as

$${}_I\omega_h(\psi(t),\dot{\psi}(t)) = \begin{bmatrix} \cos\theta & 0 \\ 0 & 1 \end{bmatrix} \cdot {}_I\begin{bmatrix} \dot{\varphi}(\psi(t),\dot{\psi}(t)) \\ \dot{\theta}(\psi(t),\dot{\psi}(t)) \end{bmatrix}. \tag{12}$$

As it is a basic principle of rigid body mechanics, this relation is not further discussed at this point.

In order to consider the current orientation of the motion system in the optimization process, the translational accelerations ${}_I a_h(t)$ and angular velocities ${}_I\omega_h(t)$ are transformed into the fixed reference system D of the driver, which is assumed to be orientated identically to the shaker reference frame (see Figure 11):

$$\begin{aligned} {}_D a_h(\psi(t),\dot{\psi}(t),\ddot{\psi}(t)) &= L_{DI} \cdot {}_I a_h(\psi(t),\dot{\psi}(t),\ddot{\psi}(t)) \\ {}_D\omega_h(\psi(t),\dot{\psi}(t)) &= T_{DI} \cdot {}_I\omega_h(\psi(t),\dot{\psi}(t)) \end{aligned} \tag{13}$$

using the rotation matrices

$$\begin{aligned} L_{DI} &= \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ \sin\varphi\cdot\sin\theta & \cos\varphi & \sin\varphi\cdot\cos\theta \\ \cos\varphi\cdot\sin\theta & -\sin\varphi & \cos\varphi\cdot\cos\theta \end{bmatrix}, \\ T_{DI} &= \begin{bmatrix} \cos\theta & 0 \\ \sin\varphi\cdot\sin\theta & \cos\varphi \end{bmatrix} \end{aligned} \tag{14}$$

At this point it becomes clear that the matrices of (12) and (14) differ from the formulations reported in literature, which is due to the fact that the ATMOS driving simulator cannot perform any rotations around the vertical axis. Consequently, the yaw angle is not taken into account, while the roll and pitch angles $\varphi(t)$ and $\theta(t)$ are determined according to (9) as functions of the state variables $\psi(t)$.

Since the low-frequency components of the longitudinal and lateral acceleration reference from the simulated vehicle cannot be replicated by translational displacements of the motion system because of its limited workspace, the previously described tilt coordination technique is applied. For this purpose, the gravitational acceleration vector g is transformed into the fixed coordinate system of the driver by means of the rotation matrix L_{DI} as

$$ {}_Dg = L_{DI} \cdot {}_I\begin{bmatrix} 0 \\ 0 \\ -g \end{bmatrix} = {}_D\begin{bmatrix} g \cdot \sin\theta(t) \\ -g \cdot \sin\varphi(t) \cdot \cos\theta(t) \\ -g \cdot \cos\varphi(t) \cdot \cos\theta(t) \end{bmatrix} . \qquad (15)$$

By combining the transformed translational accelerations of (13) with the gravitational acceleration vector ${}_Dg$ from the tilt coordination, the so-called *specific accelerations* $a(t) = {}_Da_h(t) - {}_Dg(t)$ are obtained, which are commonly used in Motion Cueing applications:

$$ a\,(t) \;=\; {}_D\begin{bmatrix} \ddot{x}\,(t) \\ \ddot{y}\,(t) \\ \ddot{z}\,(t) \end{bmatrix}_h - {}_D\begin{bmatrix} g \cdot \sin\theta(t) \\ -g \cdot \sin\varphi(t) \cdot \cos\theta(t) \\ -g \cdot \cos\varphi(t) \cdot \cos\theta(t) \end{bmatrix} \qquad (16)$$

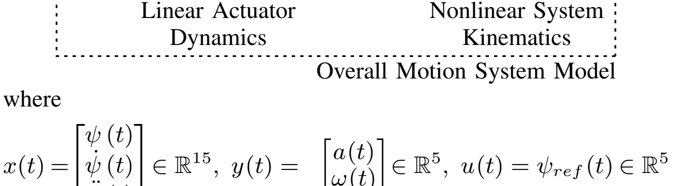
A condensed overview of the resulting process model to predict the future motion system behavior is given in Figure 12. As illustrated, it features the typical structure of a *Wiener model*, consisting of a series connection of a linear dynamic system in front of a static nonlinearity [19]. The overall system description thus is given in the form of the nonlinear state space representation

$$\begin{aligned} \dot{x}(t) &= A \cdot x(t) + B \cdot u(t) \\ y(t) &= f\,(x(t))\,. \end{aligned} \qquad (17)$$

Here, the linear state differential equation describes the dynamic transfer behavior of all controlled actuators analogously to (8). In addition, the output equation contains the kinematic relations derived in (9)–(16), summarized in the generalized vector function f, to determine the desired output variables $y(t)$ at the driver's head position within the simulator. By using the proposed model of (17), all relevant characteristics of the motion system described in Section II, such as physical limitations of the available workspace and coupling effects between individual DOF, are explicitly considered in the control algorithm, which represents one of the key features of the developed optimization-based MCA. However, the integration of all kinematic dependencies causes a significant computational effort due to the underlying model complexity. The following section therefore presents a method for efficiently calculating the future system behavior as a function of the control variables $\psi_{ref}(t)$.

B. Prediction of the Future System Behavior

In order to plan the motion trajectory of the simulator adequately for the oncoming driving situation, the future system behavior has to be specified within a limited time horizon, the so-called *prediction horizon* N, with respect to the actuating variables. This prediction is usually performed using a discrete system description, since the application of a time-continuous process model is more complex without providing any considerable benefits [20].


$u(t)$
$\dot{x}(t) = A\cdot x(t) + B\cdot u(t)$
$x(t)$
$y(t) = f\,(x(t))$
$y(t)$
Linear Actuator Dynamics
Nonlinear System Kinematics
Overall Motion System Model

where

$$ x(t) = \begin{bmatrix} \psi\,(t) \\ \dot{\psi}\,(t) \\ \ddot{\psi}\,(t) \end{bmatrix} \in \mathbb{R}^{15},\; y(t) = {}_D\begin{bmatrix} a(t) \\ \omega(t) \end{bmatrix} \in \mathbb{R}^5,\; u(t) = \psi_{ref}(t) \in \mathbb{R}^5 $$

Figure 12. Resulting Nonlinear Motion System Model.

For this reason, the solution of the state differential equation of (17) is determined using the *state-transition matrix*. According to [21] follows thus:

$$\begin{aligned} x(k+1) &= e^{A\cdot T}\cdot x(k) + \int_0^T e^{A\cdot(T-\tau)} \cdot B \cdot u(k)\,\mathrm{d}\tau \\ &= e^{A\cdot T}\cdot x(k) + \int_0^T e^{A\cdot(T-\tau)}\,\mathrm{d}\tau \cdot B \cdot u(k) \end{aligned} \qquad (18)$$

This assumes that the value of the input vector $u(k)$ does not change within the duration T of a discrete time step k, and therefore does not have to be considered within the integral. The solution of (18) then yields

$$ x(k+1) = e^{A\cdot T}\cdot x(k) + A^{-1}\cdot\left(e^{A\cdot T} - I\right)\cdot B \cdot u(k)\,. \qquad (19)$$

Here, A is required to be a nonsingular matrix, so that its inverse A^{-1} exists. For the given application, however, it can be assumed that the underlying position controls of the actuators are stable and A hence has no eigenvalues equal to zero, for which reason this condition is fulfilled here. In the following, (19) is rewritten in the more compact notation

$$ x(k+1) = A_d \cdot x(k) + B_d \cdot u(k)\,, \qquad (20)$$

with the corresponding matrices

$$\begin{aligned} A_d &= e^{A\cdot T} \\ B_d &= A^{-1}\cdot\left(e^{A\cdot T} - I\right)\cdot B. \end{aligned} \qquad (21)$$

Consequently, the time-discrete form of the state space representation (17) finally results as

$$\begin{aligned} x(k+1) &= A_d \cdot x(k) + B_d \cdot u(k) \\ y(k) &= f\,(x(k))\,. \end{aligned} \qquad (22)$$

From this, the future state variables $x(k+1) \ldots x(k+N)$ within the prediction horizon N are determined according to

$$\begin{aligned} x(k+1) &= A_d \cdot x(k) + B_d \cdot u(k) \\ x(k+2) &= A_d \cdot x(k+1) + B_d \cdot u(k+1) \\ &= A_d^2 \cdot x(k) + A_d \cdot B_d \cdot u(k) + B_d \cdot u(k+1) \\ &\;\;\vdots \\ x(k+N) &= A_d^N \cdot x(k) + A_d^{N-1} \cdot B_d \cdot u(k) \;+\; \ldots \;+ \\ &\quad\; A_d \cdot B_d \cdot u(k+N-2) + B_d \cdot u(k+N-1) \end{aligned} \qquad (23)$$

by multiplying the system matrices A_d and B_d. For the further proceeding it is recommended to formulate these expressions as a vector equation of the form:

$$ \bar{x}(k+1) = F \cdot x(k) + G \cdot \bar{u}(k) \qquad (24)$$

where

$$ \bar{x}(k+1) = \begin{bmatrix} x(k+1) \\ x(k+2) \\ \vdots \\ x(k+N) \end{bmatrix} \in \mathbb{R}^{15\cdot N},\;\; F = \begin{bmatrix} A_d \\ A_d^2 \\ \vdots \\ A_d^N \end{bmatrix} \in \mathbb{R}^{15\cdot N \times 15}, $$

$$G=\begin{bmatrix} B_d & 0 & \dots & 0 \\ A_d B_d & B_d & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_d^{N-1} B_d & A_d^{N-2} B_d & \dots & B_d \end{bmatrix} \in \mathbb{R}^{15\cdot N \times 5\cdot N},$$

$$\bar{u}(k)=\begin{bmatrix} u(k) \\ u(k+1) \\ \vdots \\ u(k+N-1) \end{bmatrix} \in \mathbb{R}^{5\cdot N} \tag{25}$$

Thus, the future state variables depend on the actual system state $x(k)$, which is known by measurement and observation, as well as the optimization variables $u(k) \dots u(k+N-1)$. Moreover, since the transfer behavior of the controlled actuators is time-invariant, the prediction matrices of (25) can already be calculated offline during the initialization process of the controller, which improves compliance with the real-time capability. In contrast, the prediction of the corresponding output variables $y(k+i)\ \forall\ i=1\dots N$, causes a large numerical effort, as these include the direct kinematic relations of the motion system. That is why an approximation of the nonlinearities of (22) within the prediction horizon is pursued in each calculation cycle, leading to a significant reduction of the computational load. Specifically, a first order *Taylor series* of the nonlinear output equation is determined as

$$y(k+i) \approx f(x(k)) + \left.\frac{\partial f(x(k))}{\partial x}\right|_{x(k)} \cdot (x(k+i) - x(k)), \tag{26}$$

where the partial derivative of the vector function f to the state vector with the value $x(k)$ yield the linear output matrix $C(k) \in \mathbb{R}^{5\times 15}$. By rearranging (26), a more structured formulation is obtained:

$$y(k+i) \approx C(k)\cdot x(k+i) + \underbrace{f(x(k)) - C(k)\cdot x(k)}_{h(k)} \tag{27}$$

Consequently, the linear affine output equation (27) results in each calculation cycle of the optimization-based controller, which approximates the nonlinear system behavior within the considered prediction horizon. Depending on the selected sampling rate, a high-frequency update of the output matrix $C(k)$ thus is performed, based on the feedback state vector $x(k)$. Furthermore, the term $h(k)$ is obtained, which depends only on the current system information and is therefore constant in the prediction range $i=1\dots N$. As this is usually limited to only a few seconds [22], the approximation of (27) provides a sufficiently accurate description of all relevant kinematic effects to optimize the simulator motion trajectory.

Although $C(k)$ and $h(k)$ must first be calculated at the beginning of each prediction sequence, the future output variables $y(k+i)$ can then be determined very efficiently according to

$$\begin{aligned} y(k+1) &= C(k)\cdot x(k+1)\ +\ h(k) \\ y(k+2) &= C(k)\cdot x(k+2)\ +\ h(k) \\ &\ \vdots \\ y(k+N) &= C(k)\cdot x(k+N) + h(k)\,. \end{aligned} \tag{28}$$

Together with the state variable prediction specified in (24), this yields the future outputs in vector notation:

$$\begin{aligned} \bar{y}(k+1) &= C\cdot \bar{x}(k+1)\ +\ H \\ &= C\cdot F\cdot x(k)\ +\ C\cdot G\cdot \bar{u}(k)\ +\ H \end{aligned} \tag{29}$$

where

$$\bar{y}(k+1)=\begin{bmatrix} y(k+1) \\ y(k+2) \\ \vdots \\ y(k+N) \end{bmatrix} \in \mathbb{R}^{5\cdot N}, \quad H=\begin{bmatrix} h(k) \\ h(k) \\ \vdots \\ h(k) \end{bmatrix} \in \mathbb{R}^{5\cdot N},$$

$$C=\begin{bmatrix} C(k) & 0 & \dots & 0 \\ 0 & C(k) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & C(k) \end{bmatrix} \in \mathbb{R}^{5\cdot N \times 15\cdot N} \tag{30}$$

As a result, (29) offers the advantage that only its first two summands have to be evaluated at runtime by simple matrix multiplications, instead of evaluating the nonlinear output equation of (22) for each single time step $k+i$ within the prediction horizon $i=1\dots N$.

C. Solution of the Optimal Control Problem

In order to reproduce the vestibular Motion Cues of the simulated vehicle, given by its translational accelerations and angular velocities, the optimal control problem

$$\begin{aligned} \underset{\Delta\bar{u}(k)}{\text{minimize}} \sum_{i=1}^{N} \| y\,(k+i) - r\,(k+i)\,\|_Q^2 \ + \sum_{i=1}^{N} \rho\,(k+i) \\ + \sum_{i=0}^{N-1} \| \Delta u\,(k+i)\,\|_R^2 \ +\ \| u\,(k+N-1)\,\|_S^2 \end{aligned} \tag{31}$$

subject to

$$\begin{aligned} x_{lo} \leq x\,(k+i) \leq x_{up},\ \ i \in [1, N] \\ u_{lo} \leq u\,(k+i) \leq u_{up},\ \ i \in [0, N{-}1] \end{aligned}$$

is solved numerically in each calculation cycle of the MPC-based algorithm. Here, the first and third summand of the cost function evaluate the control deviation as well as the change rate of the actuating variables $\Delta u(k) \dots \Delta u(k+N-1)$ for all time steps in prediction horizon, using the positive definite weighting matrices $Q \in \mathbb{R}^{5\times 5}$ and $R \in \mathbb{R}^{5\times 5}$. The control deviation results from the difference between the future output variables, which are expressed according to (29) as a function of the feedback state vector $x(k)$ and the optimization variables, and the simulated vehicle accelerations and angular velocities summarized in the reference vector

$$r(k+i) = \begin{bmatrix} a_{ref}(k+i) \\ \omega_{ref}(k+i) \end{bmatrix} \in \mathbb{R}^5 \ \ \forall\ i=1\dots N. \tag{32}$$

However, since these references depend on the future driver inputs in the prediction horizon, they are generally not exactly known in the current time step k. It is therefore common practice to consider the vehicle references constant at each future time step, although this does not fully exploit the potential of the predictive controller [23]. To overcome this, we proposed a novel model-based online prediction strategy in [24]. As key features, this approach includes a simplified vehicle model as well as a virtual driver model based on established algorithms from nonlinear control theory to estimate future driver inputs and the resulting vehicle reactions depending on the current driving situation and given route information. By means of measurement data from a real test drive, it was proven that the reproduced Motion Cues differ only slightly from those of an exactly known reference trajectory, which demonstrates the effectiveness of the developed approach. In the context of this paper, however, an a-priori known future reference is assumed, since the general functionality of the predictive MCA and its handling of the considered motion systems kinematic properties are to be highlighted.

In addition, the second summand of (31) denotes a *penalty term* to prevent deviations between the angular velocities of the reference signal and those of the motion system above a defined boundary. This enables the tilt coordination rotation rate as well as the forced rotations due to the kinematic couplings of the motion platform to be limited to a desired value ε, for example the perception threshold of the human vestibular organs:

$$\rho\,(k+i) = e^{\sigma\cdot(|\omega(k+i)-\omega_{ref}(k+i)|-\varepsilon)} \tag{33}$$

Selecting appropriate penalty weights $\sigma \gg 1$, the limitation of the rotation rate is taken into account in the numerical optimization, since the penalty term applies:

$$\rho\,(k+i)\begin{cases} \ll 1 & \text{if } |\,\omega(k+i)-\omega_{ref}(k+i)\,| < \varepsilon \\ = 1 & \text{if } |\,\omega(k+i)-\omega_{ref}(k+i)\,| = \varepsilon \\ \gg 1 & \text{if } |\,\omega(k+i)-\omega_{ref}(k+i)\,| > \varepsilon \end{cases} \tag{34}$$

Furthermore, the last element of the cost function represents a terminal cost to create a washout effect and return the simulator to its initial position. Here, the positive definite weighting matrix $S \in \mathbb{R}^{5\times5}$ determines the intensity of the washout movement. To comply with the physical limitations of the motion system, constraints on the state and actuating variables are included in (31). For this, lower and upper boundaries $(\cdot)_{lo}$ and $(\cdot)_{up}$ are defined according to the installed actuators performances and the available workspace.

The resulting optimal control problem is solved at runtime on the dSPACE DS1006 system with a sampling time of $25\,ms$, using the *conservative convex separable approximation* (CCSA) algorithm [25], which is provided by the *NLopt* open-source library for nonlinear optimization [26]. Thereby, the prediction horizon is chosen to $N = 40$ discrete time steps to realize a receding time horizon of one second. The constrained optimal control problem of (31) hence involves 200 optimization variables $\Delta\bar{u}(k)$ that are determined in real-time by the proposed control strategy. Figure 13 schematically shows the overall signal processing structure in a block diagram. In addition to the previously described methodology, it includes a state observer [27], enabling the complete state vector $x(k)$ to be determined in each time step k from the measured angular positions $\psi(k)$ and velocities $\dot{\psi}(k)$. The basis of this observer are the dynamic models of the controlled actuators according to (4).

V. Comparison of the Control Strategies

Since the scientific objective of this paper deals with the comparison of the filter- and optimization-based control algorithms presented in Sections III and IV, the underlying evaluation framework is described in detail at this point. The applied quality criteria are initially discussed for this purpose. Afterwards, the driving scenarios examined in this study will be briefly introduced.

A. Objective Quality Criteria

In order to compare both Motion Cueing strategies on the basis of an objective valuation metric, suitable quality criteria must be specified. Therefore, according to [28] and [29], we introduce performance indicators λ_1 and λ_2 that are defined as

$$\begin{aligned} \lambda_1 = \frac{1}{M}\sum_{j=0}^{M}\sqrt{\left(\frac{e_{a_x,j}}{a_{x,norm}}\right)^2+\left(\frac{e_{a_y,j}}{a_{y,norm}}\right)^2+\left(\frac{e_{a_z,j}}{a_{z,norm}}\right)^2} \\ +\frac{1}{M}\sum_{j=0}^{M}\sqrt{\left(\frac{e_{\omega_x,j}}{\omega_{x,norm}}\right)^2+\left(\frac{e_{\omega_y,j}}{\omega_{y,norm}}\right)^2} \end{aligned} \tag{35}$$

and

$$\begin{aligned} \lambda_2 = \frac{1}{M}\sum_{j=0}^{M}\sqrt{\left(\frac{e_{\hat{a}_x,j}}{a_{x,norm}}\right)^2+\left(\frac{e_{\hat{a}_y,j}}{a_{y,norm}}\right)^2+\left(\frac{e_{\hat{a}_z,j}}{a_{z,norm}}\right)^2} \\ +\frac{1}{M}\sum_{j=0}^{M}\sqrt{\left(\frac{e_{\hat{\omega}_x,j}}{\omega_{x,norm}}\right)^2+\left(\frac{e_{\hat{\omega}_y,j}}{\omega_{y,norm}}\right)^2} \end{aligned} \tag{36}$$

with

$$\begin{aligned} e_{a_i} &= a_{i,Ref} - a_i|_{i=x,y,z} \text{ and } e_{\omega_i} = \omega_{i,Ref} - \omega_i|_{i=x,y} \\ e_{\hat{a}_i} &= \hat{a}_{i,Ref} - \hat{a}_i|_{i=x,y,z} \text{ and } e_{\hat{\omega}_i} = \hat{\omega}_{i,Ref} - \hat{\omega}_i|_{i=x,y}. \end{aligned} \tag{37}$$

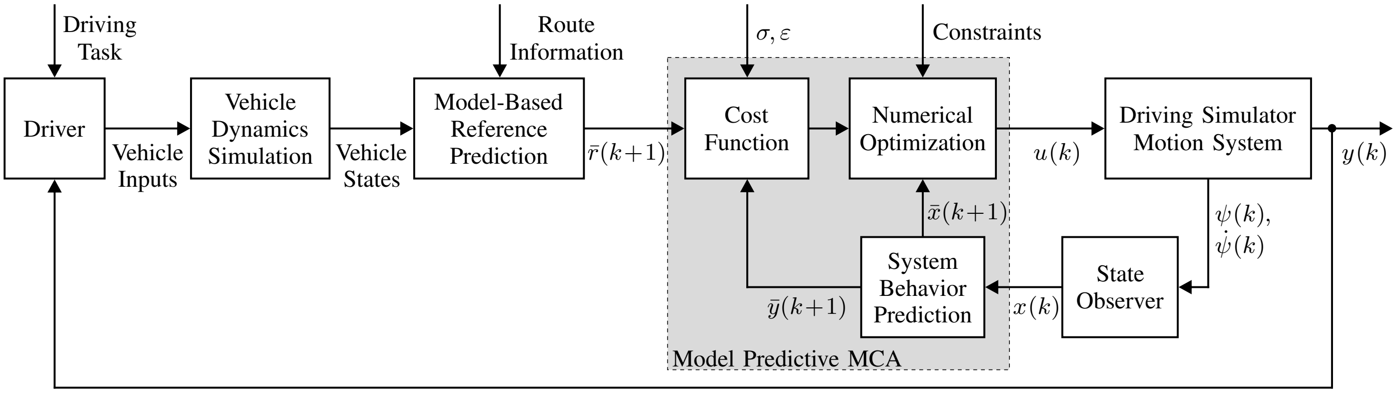


Figure 13. Signal Processing Structure of the MPC-Based Motion Cueing Strategy.

Here, (35) provides a measure of the physical deviations between the scaled reference accelerations $a_{i,Ref}$ and angular velocities $\omega_{i,Ref}$ from the vehicle dynamics simulation and the measured quantities in the driving simulator for the considered DOF. λ_1 therefore returns the averaged normalized control error over the number of measured values M within the considered time range. The normalization is necessary to obtain dimensionless quantities that allow a simultaneous consideration of accelerations and angular velocities on a common scale. According to [30], the human perception thresholds for movements are used as corresponding normalization factors $a_{i,norm}$ and $\omega_{i,norm}$. In addition, the indicator λ_2 as defined by (36) yields a measure for the perceived control quality, which can differ from the physical deviations due to the frequency-dependent dynamic behavior of the human vestibular organs, as well as perception thresholds. This causes, for example, that control errors in detectable frequency ranges are perceived more disturbing than deviations in undetectable ranges. To take these effects into account, well-established models of the human vestibular system illustrated in Figure 14 are included. Here, the primary perceptual organs are the semicircular canals, which enable the detection of angular velocities in all three rotational DOF, and the otoliths that are responsible for the perception of longitudinal, lateral and vertical accelerations.

According to Figure 15, the corresponding dynamic behavior is typically described by mechanical analogous models of the respective organs, which lead to the illustrated transfer functions with the inputs a_i and ω_i [32][33], as they are widely used in driving simulation applications [14]. In agreement with [34], the parameters of the otoliths model are selected to $K_{oto} = 0.4$, $T_1 = 5\,s$, $T_2 = 0.016\,s$ and $T_L = 10\,s$, while the semicircular canal model parameters are $K_{scc} = 5.73$, $T_1 = 5.73\,s$, $T_2 = 0.005\,s$, $T_L = 0.06\,s$ and $T_a = 80\,s$. This leads to the resulting frequency responses of the transfer functions $G_{oto}(j\omega)$ and $G_{scc}(j\omega)$ shown in Figure 16. It becomes clear that the semicircular canals serve as good angular velocity sensors in the frequency range from 0.05 to $3\,Hz$, since rotary motions are closely detected without amplitude changes and with only small phase shifts. This frequency spectrum is also characteristic for everyday driving maneuvers in traffic, which is why rotary vehicle movements can be easily perceived by the human vestibular apparatus. In contrast, low-frequency rotations are perceived strongly damped and are almost completely suppressed in case of a constant angular velocity. These characteristics of the semicircular canals are used in interactive driving simulation to apply the previously described tilt coordination technique without the driver being able to detect the unnatural rotational movements. In addition, the modeled otoliths show a frequency-specific filter behavior. Analogous to the semicircular canals model, the passband is found at frequencies of 0.05 to $3\,Hz$, in which the perceived accelerations $\bar{a}$ at the transfer function outputs contain only a slight amplitude attenuation and phase shift. Thus, the otoliths provide very good acceleration sensors in the frequency range of common driving maneuvers so that translational vehicle movements can be precisely detected. However, low-frequency acceleration stimuli below $0.05\,Hz$ are only perceived with an amplitude attenuated by about $-8\,dB$. In the high-frequency range, a characteristic low-pass behavior is observed, which is due to the inertia of the otoliths. As a consequence, accelerations above $20\,Hz$, e.g., high-frequency engine vibrations, are only sensed inaccurately by the vestibular organs, so that further perception systems are required for a correct interpretation of the actual motion.

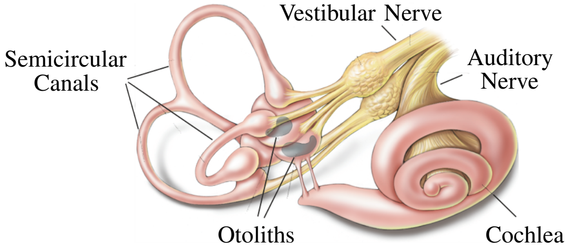


Figure 14. Vestibular System in the Human Inner Ear [31].

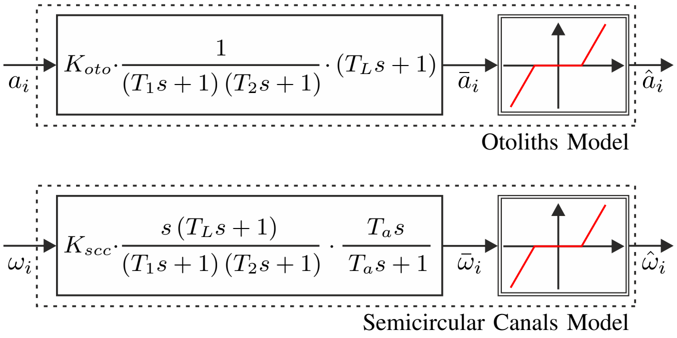

a_i
$K_{oto} \cdot \frac{1}{(T_1 s + 1)(T_2 s + 1)} \cdot (T_L s + 1)$
$\bar{a}_i$
$\hat{a}_i$
Otoliths Model
ω_i
$K_{scc} \cdot \frac{s(T_L s + 1)}{(T_1 s + 1)(T_2 s + 1)} \cdot \frac{T_a s}{T_a s + 1}$
$\bar{\omega}_i$
$\hat{\omega}_i$
Semicircular Canals Model

Figure 15. Applied Models of the Vestibular Organs.

By a series connection of the transfer functions with nonlinear dead zones (see Figure 15), the threshold values $a_{i,thres}$ and $\omega_{i,thres}$ of the human perception are integrated

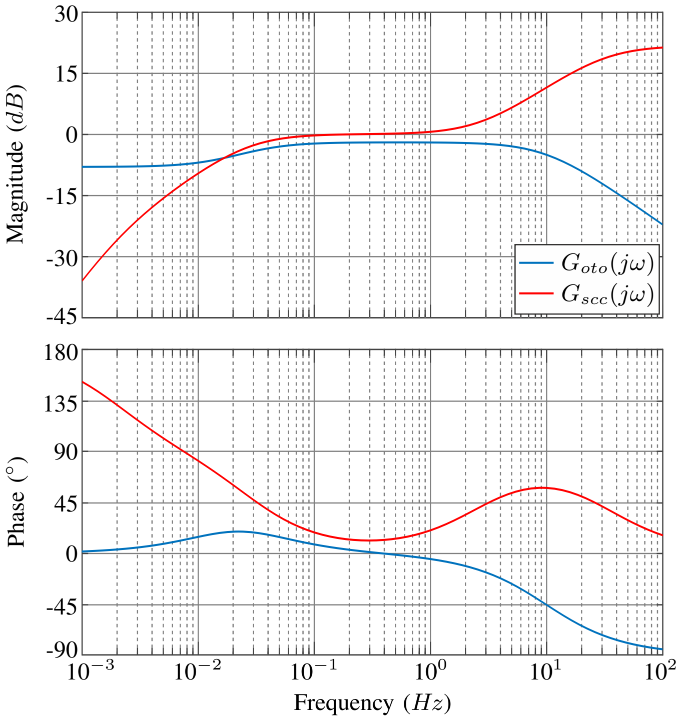


Figure 16. Frequency Responses of the Applied Transfer Function Models.

with respect to the following relationship [8]:

$$\hat{a}_i = \begin{cases} 0 & \text{if } |\bar{a}_i| \leq a_{i,thres} \\ \bar{a}_i - sgn\left(\bar{a}_i\right) \cdot a_{i,thres} & \text{if } |\bar{a}_i| > a_{i,thres} \end{cases}$$
$$\hat{\omega}_i = \begin{cases} 0 & \text{if } |\bar{\omega}_i| \leq \omega_{i,thres} \\ \bar{\omega}_i - sgn\left(\bar{\omega}_i\right) \cdot \omega_{i,thres} & \text{if } |\bar{\omega}_i| > \omega_{i,thres} \end{cases} \quad (38)$$

Consequently, the closer the performance indicators λ_1 and λ_2 are to the origin, the better is the reproduction of the simulated vehicle movements, whereby the value zero indicates a perfect motion rendering. However, especially with regard to λ_1, this is only a theoretical value that cannot be obtained by any driving simulator, since it would require an almost unlimited workspace.

B. Driving Scenarios

For the purpose of obtaining a representative comparison of the two control strategies, a selection of nine driving scenarios was defined. These contain standardized maneuvers, which are commonly used for development and optimization applications in the automotive industry, like:

- Acceleration from standstill
- Braking from driving straight forward (DIN ISO 70028)
- Lane change (DIN ISO 3888-1)
- Step steering (DIN ISO 7401)
- Braking from steady-state circular course drive (DIN ISO 7975)

As the listed maneuvers are mainly used to identify and analyze the driving dynamics of a vehicle, they do not represent usual driving situations. For this reason, also moderate scenarios are examined in the evaluation:

- Turning at a junction
- Drive on a rural road
- Drive through a roundabout
- Drive through a highway interchange

Vehicle dynamics simulations of all nine maneuvers were performed and the relevant accelerations and angular velocities were recorded. Subsequently, these data were used as identical reference signals for both MCA to ensure a consistent basis for evaluation described in the next section.

VI. Results and Discussion

Subsequently, the results of the comparison of the two Motion Cueing strategies are presented and the impacts on the interactive driving simulation are discussed. For that purpose, both control algorithms were implemented on the ATMOS driving simulator. Measurement data of the translational accelerations and the angular velocities taken with the installed IMU at the driver's head position serve as inputs for the quality criteria presented in Section V. For reasons of clarity, only the measured data of one driving scenario from each maneuver class are analyzed in detail. All further scenarios will be summarized in the following.

A. Scenario Acceleration from Standstill

First the maneuver "acceleration from standstill" is discussed, in which the simulated vehicle accelerates from standstill to a given speed of $130\,km/h$. Thereby no steering movements of the driver take place, so that there is no lateral vehicle excitation. Figure 17 shows the resulting longitudinal acceleration and pitch velocity tracking using both MCA. It becomes clear that an adequate reproduction quality of the longitudinal acceleration from the vehicle dynamics simulation is achieved regardless of the applied algorithm. Only when the reference rises rapidly at time $t = 4\,s$, there are significant deviations between the simulated and measured acceleration in the driving simulator. In case of the washout algorithm, these can be explained by the signal processing of the washout filters that are used to move the motion system back to the neutral position. At the same time, the tilt coordination rotation is restricted to the delayed dynamics of the low-pass filters, resulting in the illustrated control error. The MPC approach, in contrast, achieves notably smaller deviations. Nevertheless, even with this algorithm, the simulated vehicle acceleration cannot be reproduced exactly, which can be attributed to the limited pitch velocity. As explained in Section IV, the overall rotation rate error of the motion system is bounded to the value of $\varepsilon = 0.1\,rad/s$ so that unexpected rotations caused by the tilt coordination technique and the kinematic couplings of the motion platform are not perceived disturbingly by the driver [35]. Thus, acceleration deviations, as shown at time $t = 4\,s$, are allowed by the optimization algorithm to keep the rotations of the motion system below the perception threshold of the vestibular organs. Without this rotation rate limitation or when using a motion system without couplings between translational and rotational DOF, such as a hexapod, the simulated vehicles acceleration could be reproduced almost exactly in the simulator. In addition, the measured pitch

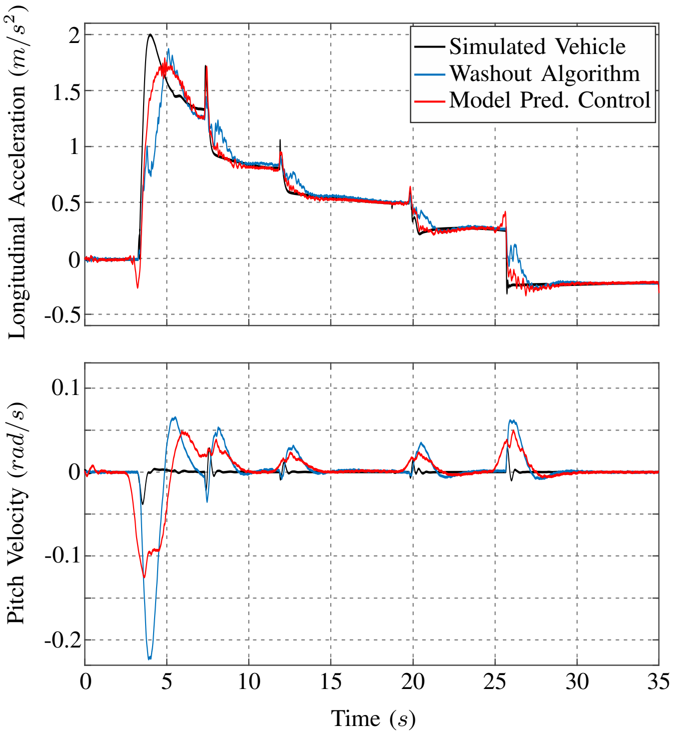


Figure 17. Longitudinal Acceleration and Pitch Velocity Tracking.

velocity in Figure 17 contains in both cases low-frequency disturbances to the vehicle reference resulting from the tilt coordination technique and the forced rotations of the motion platform. When using the filter-based MCA, these deviations are approximately twice as large at the moment of acceleration increase as with the model predictive controller, so that it is to be expected that they have a negative impact on the resulting driving impression. In Figure 18 the lateral acceleration and the corresponding roll velocity tracking are illustrated. As there are no steering actions in this maneuver, the reference values are zero throughout the observed time range. Accordingly, the measured accelerations also provide values close to zero, with only minor deviations due to measurement inaccuracies. However, these are far below the perception threshold and are therefore not noticeable for the driver. Since each translational movement of the motion platform simultaneously causes a vertical displacement of the platform center point, the measured accelerations in Figure 19 contain unpreventable low-frequency errors compared to simulated vehicle acceleration. Due to the available model knowledge, the optimization-based MCA plans the motion trajectory of the simulator in such a way that these deviations are kept below the perception threshold of the otoliths. Furthermore, additionally acting vertical acceleration references, such as at time $t = 26\,s$, are reproduced with high control quality. On the other hand, the washout algorithm generates clearly higher vertical accelerations, since like in the Classical Washout Algorithm, the translational degrees of freedom are considered independently of each other in this approach. Based on these measurement results, the application of the introduced quality criteria provides performance indicators of $\lambda_{1,WO} = 0.68$ and $\lambda_{2,WO} = 0.35$ for the washout algorithm and $\lambda_{1,MPC} = 0.48$ as well as $\lambda_{2,MPC} = 0.18$ for the predictive controller. This objectification confirms the assumption that a higher quality of motion rendering can be achieved using the optimization-based MCA as smaller performance indicators are obtained. An explanation for these results can be found in a more efficient coordination of the motion platform and the shaker system by the MPC. To illustrate this in more detail, Figure 20 shows the actuating variables determined by both approaches during the experiment. Here it can be seen that the actuator reference angles ψ_{ref} in the longitudinal and lateral direction of the motion platform as well as the three shaker actuators located on the left, the right and at the rear remain always within the simulator workspace limitations. However, also the generally different functioning of the two control strategies becomes

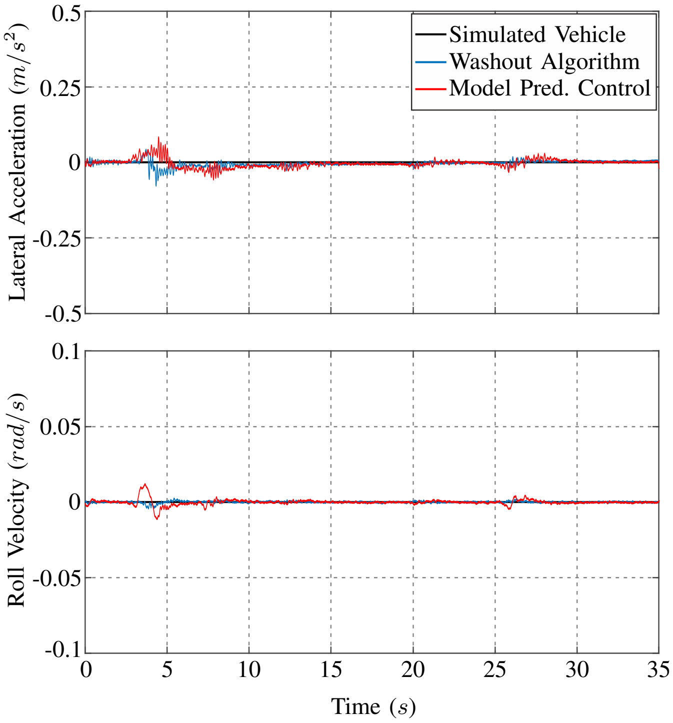


Figure 18. Lateral Acceleration and Roll Velocity Tracking.

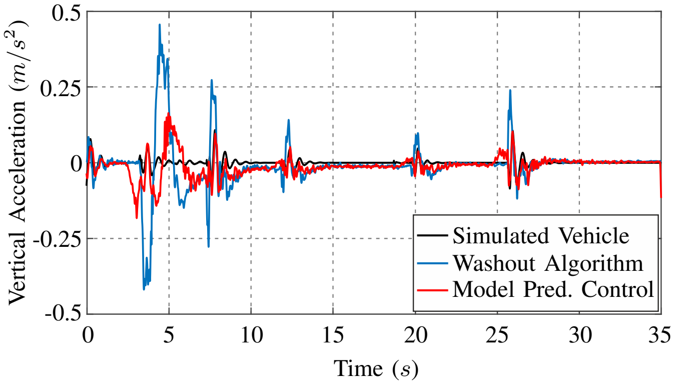


Figure 19. Vertical Acceleration Tracking.

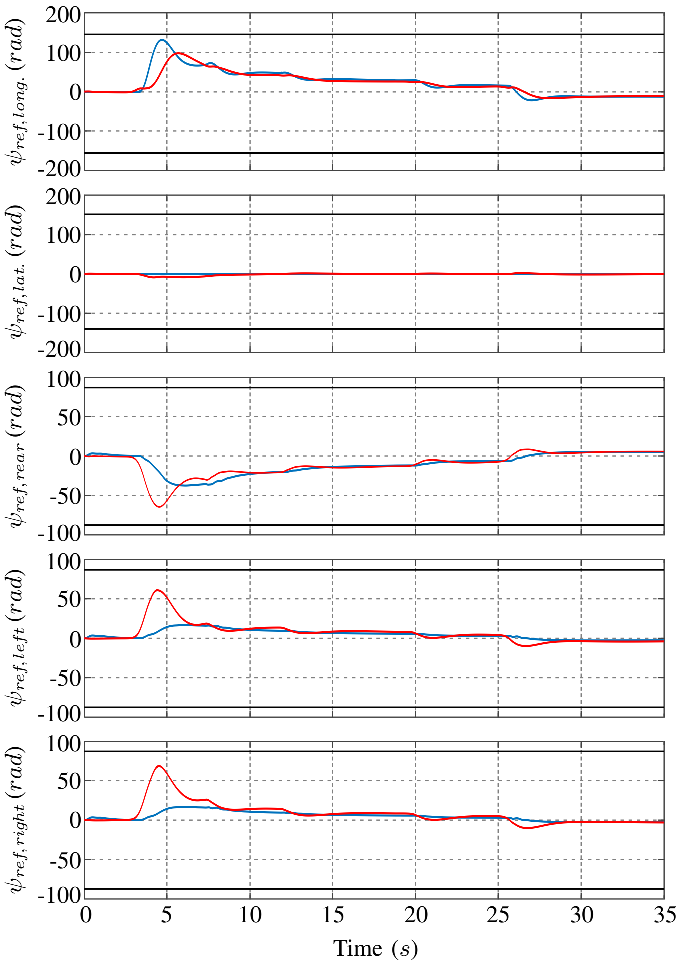


Figure 20. Comparison of the Actuating Variables: Limitation (–), Washout Algorithm (–), Model Predictive Control (–).

clear. While the coordination between motion platform and shaker in the filter-based algorithm is mainly predefined via the static filter parameters and the distribution coefficients, both subsystems are controlled by the MPC according to the current driving situation and the actual state of the motion system. For this reason, there is a variable distribution between motion platform and shaker in each driving scenario. Both systems are thereby used asynchronously in order not to exceed the rotation rate limitation due to the coupled DOF and the nonlinear kinematics of the motion system, as can be observed at time $t = 4\,s$. In addition, there is a better exploitation of the available workspace by the model predictive control algorithm.

B. Scenario Turning at a Junction

As an example of an everyday driving situation, the scenario “turning at a junction” with simultaneously acting longitudinal and lateral acceleration references will be examined subsequently. In contrast to the previously discussed maneuver, the reproduction of lateral accelerations using the presented Motion Cueing strategies can thus also be analyzed. Figure 21 illustrates the tracking of the simulated vehicles longitudinal acceleration and pitch velocity. Again it becomes clear that both the washout algorithm and the optimization-based MCA yield an adequate reproduction of the longitudinal acceleration. However, the measured accelerations show, such as at time $t = 10\,s$, a larger delay in comparison to the reference signal when using the washout algorithm due to the phase shift of the implemented filters. Also in this maneuver, the associated pitch velocity contains in both cases low-frequency disturbances that can be explained by the tilt coordination, since sustained acceleration components can only be reproduced by an equivalent rotation of the motion system. Using the washout algorithm, these errors are significantly higher due to the forced rotation of the motion platform, so it can be expected that the resulting driving experience will be negatively affected. In contrast, the predictive MCA uses the integrated kinematics information to successfully limit the overall rotation rate error to $0.1\,rad/s$. As a result of this limitation, minor errors in the tracking of the acceleration reference occur, which are more difficult to detect by the driver in the simulator than unexpected strong rotations. Equivalent results can be derived from Figure 22, that illustrates the lateral acceleration and the corresponding roll velocity. As shown, the acceleration reference from the vehicle dynamics simulation is tracked very well with both algorithms. There are again time delays to the reference signal that are larger when using the washout algorithm, resulting from the phase shift of the implemented filters. The roll velocity error is also larger compared to the MPC, even if the difference between both algorithms is smaller than in case of the pitch velocity. Thus, as a consequence for the interactive driving simulation, the resulting driving experience can be expected to be more realistic using the predictive control strategy, since smaller rotation rate errors are more difficult to detect for the human perception system. The vertical acceleration measured in the examined driving scenario is illustrated in Figure 23. Also in this maneuver it is noticeable that due to the coupled DOF of the motion system, undesired vertical displacements occur, which cannot be fully compensated by either control strategy. However, these errors are significantly lower and mostly below the human perception threshold in the use of the predictive MCA. The washout algorithm, on the other hand, generates detectable sensory conflicts since no interactions between horizontal and vertical accelerations are considered in the underlying algorithm structure. To objectify these findings, the quality criteria introduced in the previous section are used, resulting in performance indicators $\lambda_{1,WO} = 1.74$ and $\lambda_{2,WO} = 0.92$ for the washout algorithm and $\lambda_{1,MPC} = 1.20$ and $\lambda_{2,MPC} = 0.53$ for the optimization-based MCA. It becomes consequently clear that a higher control quality is achieved with the MPC, which is primarily explained by the

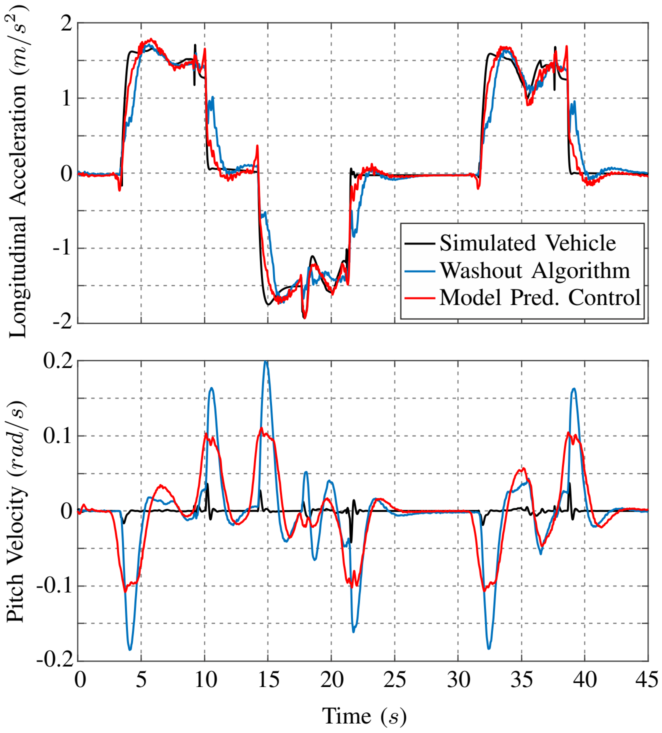


Figure 21. Longitudinal Acceleration and Pitch Velocity Tracking.

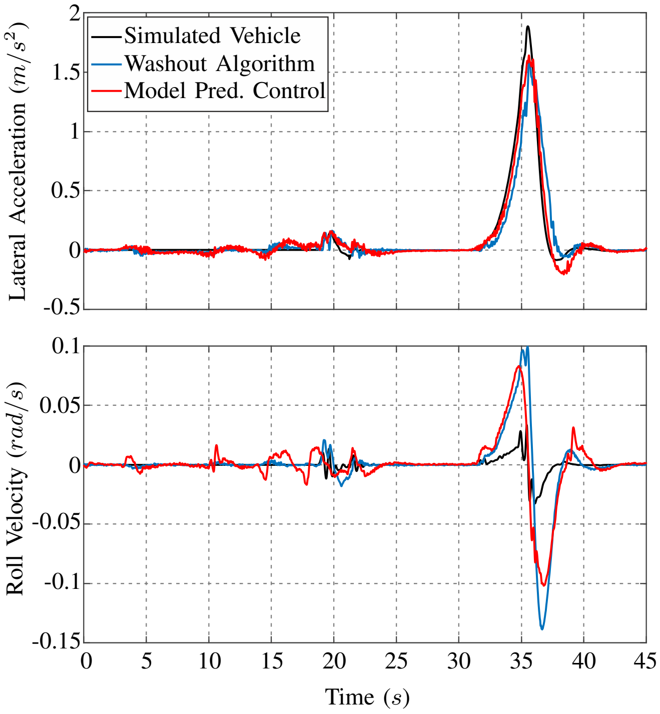


Figure 22. Lateral Acceleration and Roll Velocity Tracking.

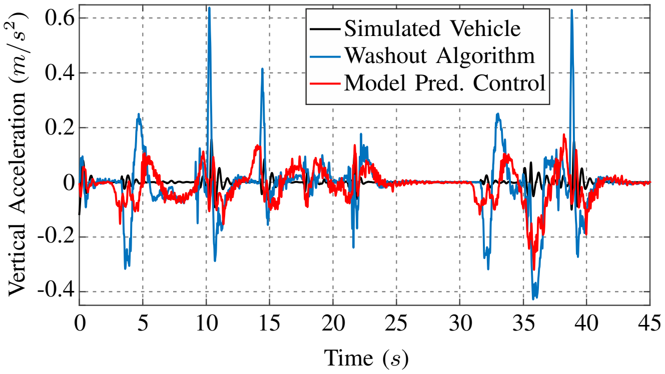


Figure 23. Vertical Acceleration Tracking.

lower angular velocity and vertical acceleration errors caused by the specific kinematics of the ATMOS driving simulator. Here, the differences between filter-based and optimization-based MCA are again obvious when considering the associated actuating variables in Figure 24. Although both algorithms respect the available workspace of the installed actuators at all times, the coordination of the motion platform and the shaker system shows significant differences. Similar to the example of the previously considered driving scenario, the shaker is used more in the model predictive algorithm in order to compensate the coupling effects of the motion platform best possible. Thereby, the motion trajectories of both subsystems are planned asynchronously (see e.g., at time $t = 10\,s$) to comply with the given rotation rate limitations of $0.1\,rad/s$ while reproducing the acceleration references from the simulated vehicle. In the washout algorithm, in contrast, there are no compensation operations with the shaker, resulting in the rotation rate and vertical acceleration errors illustrated in the Figures 21, 22 and 23.

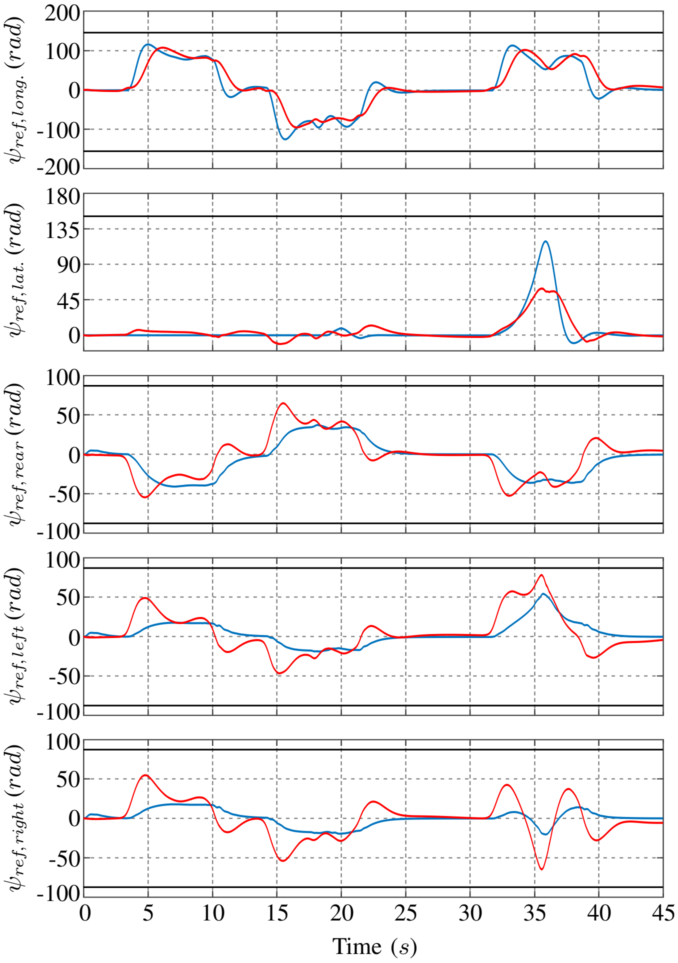


Figure 24. Comparison of the Actuating Variables: Limitation (–), Washout Algorithm (–), Model Predictive Control (–).

C. Summarized Evaluation of all Driving Scenarios

The evaluation process described before using the example of two selected driving maneuvers was performed for all nine test scenarios in the context of this paper. Thereby, the performance indicators listed in Table II were obtained. A graphical analysis of these results can be seen in Figure 25, which combines the evaluation of all maneuvers in a common radar chart. Here, the two driving scenarios "acceleration from standstill" and "turning at a junction" exhibit the lowest and the highest performance indicators respectively. But it should be noted that the individual maneuvers are not comparable with each other, as they differ significantly in terms of the underlying driving dynamics. For example, purely longitudinal scenarios such as "braking from driving straight forward" naturally generate lower values of λ_1 and λ_2 than more challenging maneuvers with simultaneously acting longitudinal and lateral accelerations. However, the presented evaluation framework enables a reliable objective comparison of both Motion Cueing strategies for each separate driving scenario. The chart clearly shows the advantages of the optimization-based MCA in comparison to the washout algorithm, since smaller performance indicators are achieved in each of the examined scenarios. Here, it is noticeable that the perceived control quality, expressed by the indicator λ_2, yields small values close to zero when the MPC is used and therefore a good subjective driving impression can be expected. As already discussed in detail in the previous sections, these results can be explained with the angular velocity and vertical acceleration errors due to the coupled degrees of freedom, because of which an adequate reproduction of the simulated vehicles Motion Cues is a challenging task. Here, it is a great advantage of the MPC that the specific simulator kinematics are directly considered via existing model knowledge in the optimization algorithm. This allows undesired interactions to be taken into account in the planning of the motion trajectory and optimally compensated according to the current driving situation, which is a major benefit for interactive driving simulation.

TABLE II. DETERMINED PERFORMANCE INDICATORS.

Driving Scenario	$\lambda_{1,WO}$	$\lambda_{2,WO}$	$\lambda_{1,MPC}$	$\lambda_{2,MPC}$
Acceleration from Standstill	0.68	0.35	0.48	0.18
Braking from Driving Straight Forward	0.53	0.25	0.39	0.14
Lane Change	1.77	0.99	1.12	0.51
Step Steering	1.38	0.98	0.67	0.36
Braking from Steady-State Circular Course Drive	0.91	0.40	0.62	0.20
Turning at Junction	1.74	0.92	1.20	0.53
Drive Through Rural Road	1.19	0.60	0.81	0.30
Drive Through Roundabout	1.47	0.80	0.96	0.41
Drive Through Highway Interchange	0.96	0.42	0.58	0.19

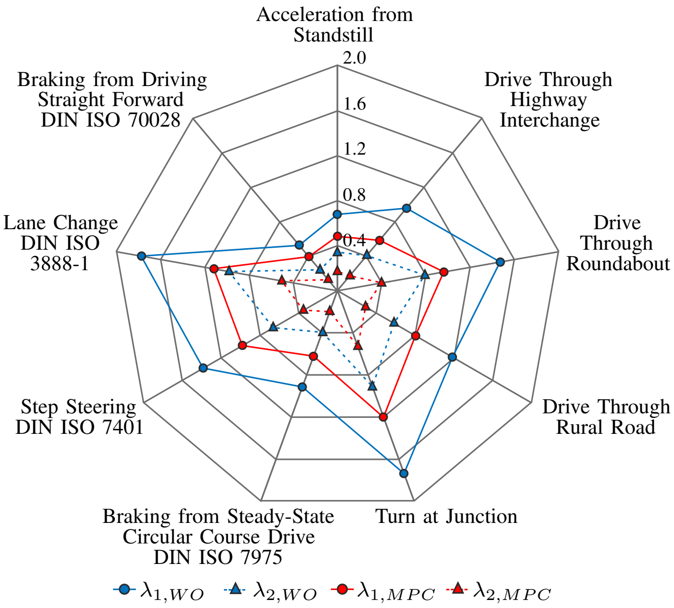


Figure 25. Evaluation of the Analyzed Test Maneuvers.

VII. Conclusion and Future Work

In this paper, the development of different Motion Cueing Algorithms for a hybrid kinematics driving simulator with 5 degrees of freedom was presented. Motivated by the unique characteristics of the considered motion system, a comprehensive extension of the filter-based Classical Washout Algorithm was designed first. Key features of the resulting control strategy form a dynamic position washout to any point within the simulator workspace, as well as a tilt coordination distribution strategy in order to make full use of the motion capabilities. However, similar to the basic idea of the CWA, this approach does not consider couplings between the individual translational DOF, which leads to undesired interdependencies that may disturb the driving impression under certain circumstances. To overcome this, an optimization-based MCA using the concept of Model Predictive Control was implemented. It includes a simplified model of the controlled actuators as well as the nonlinear kinematic relations of the motion system to optimally plan the trajectory of the simulator in real-time, taking into account given constraints. Thus, the physical limits of the system, such as the restricted workspace, are respected and the occurring coupling effects are compensated best possible.

To analyze the resulting control quality, both algorithms were objectively compared by means of defined quality criteria and standard driving scenarios from the automotive industry. Thereby, a satisfactory motion rendering was proven for each Motion Cueing strategy. However, due to the integration of model knowledge, the predictive MCA exhibits less control errors in angular velocities and vertical acceleration. For this reason, it is assumed that the subjective driving impression is more realistic when using the MPC, which is why this approach offers great potential for interactive driving simulation. On the other hand, the filter-based MCA has the advantages of simple implementation, good traceability and low computational effort, which relativizes the worse control quality in comparison to the optimization-based algorithm.

The future work will deal with the subjective validation of our observations. In this context, reliable subject studies will be conducted in order to rate the resulting degree of immersion by human Drivers-in-the-Loop. Thus, it will be possible to investigate by paired comparison of both approaches whether there is a correlation between the perceived control performance and the objective results presented in this paper. In addition, methods from the field of decoupling control theory can be integrated in the modified washout algorithm to compensate the vertical movements of the motion platform with the shaker in a limited area of the workspace, so that occurring False Cues are reduced more effectively.

References

[1] P. Biemelt, S. Mertin, N. Rüddenklau, S. Gausemeier, and A. Trächtler, “Objective Evaluation of a Novel Filter-Based Motion Cueing Algorithm in Comparison to Optimization-Based Control in Interactive Driving Simulation,” 11th International Conference on Advances in System Simulation (SIMUL), 2019, pp. 25–31.

[2] N. Rüddenklau, P. Biemelt, S. Henning, S. Gausemeier, and A. Trächtler, “Real-Time Lighting of High-Definition Headlamps for Night Driving Simulation,” International Journal On Advances in Systems and Measurements, vol. 12, 2019, pp. 72–88.

[3] V. Melcher, S. Rauh, F. Diederichs, H. Widlroither, and W. Bauer, “Take-Over Requests for automated driving,” Procedia Manufacturing, vol. 3, 2015, pp. 2867–2873.

[4] H. Bellem et al., “Can We Study Autonomous Driving Comfort in Moving-Base Driving Simulators? A Validation Study,” Human Factors, vol. 59, no. 3, 2017, pp. 442–456.

[5] T. Murano, T. Yonekawa, M. Aga, and S. Nagiri, “Development of High-Performance Driving Simulator,” SAE International Journal of Passenger Cars-Mechanical Systems, vol. 2, 2009, pp. 661–669.

[6] E. Zeeb, “Daimler’s New Full-Scale, High-dynamic Driving Simulator – A Technical Overview,” Driving Simulation Conference Europe (DSC), 2010, pp. 157–165.

[7] S. F. Schmidt and B. Conrad, “Motion Drive Signals for Piloted Flight Simulators,” I Contract Report NASA, CR-1601, 1970.

[8] L. D. Reid and M. A. Nahon, “Flight Simulation Motion-Base Drive Algorithms: Part 1 - Developing and Testing the Equations,” University of Toronto, UTIAS Report 296, 1985.

[9] T. Sammet, “Motion-Cueing-Algorithmen für die Fahrsimulation (Motion Cueing Algorithms for Driving Simulation),” Fortschritt Berichte-VDI Reihe 12 Verkehrstechnik/Fahrzeugtechnik, vol. 643, 2007.

[10] K. Patil, S. K. Molla, and T. Schulze, “Hybrid Vehicle Model Development using ASM-AMESim-Simscape Co-Simulation for Real-Time HIL Applications,” SAE Technical Paper, Tech. Rep., 2012.

[11] D. Stewart, “A Platform with Six Degrees of Freedom,” Proceedings of the Institution of Mechanical Engineers, vol. 180, no. 1, 1965, pp. 371–386.

[12] G. Bertolini and D. Straumann, “Moving in a Moving World: A Review on Vestibular Motion Sickness,” Frontiers in Neurology, vol. 7, no. 14, 2016.

[13] P. Biemelt, S. Mertin, N. Rüddenklau, S. Gausemeier, and A. Trächtler, “Design and Evaluation of a Novel Filter-Based Motion Cueing Strategy for a Hybrid Kinematics Driving Simulator with 5 Degrees of Freedom,” Driving Simulation Conference Europe (DSC), 2020.

[14] M. Bruschetta, F. Maran, and A. Beghi, “A fast implementation of MPC-based motion cueing algorithms for mid-size road vehicle motion simulators.” Vehicle System Dynamics, vol. 55, no. 6, 2017, pp. 802–826.

[15] J. O. Pitz, “Vorausschauender Motion-Cueing-Algorithmus für den Stuttgarter Fahrsimulator (Predictive Motion Cueing Algorithm for the Stuttgart Driving Simulator),” PhD Thesis, Universität Stuttgart, 2017.

[16] P. Biemelt, S. Henning, N. Rüddenklau, S. Gausemeier, and A. Trächtler, “A Model Predictive Motion Cueing Strategy for a 5-Degree-of-Freedom Driving Simulator with Hybrid Kinematics,” Driving Simulation Conference Europe (DSC), 2018, pp. 79–85.

[17] J. Richalet and D. O’Donovan, Predictive Functional Control: Principles and Industrial Applications. Springer Science & Business Media, 2009.

[18] A. J. Benson, E. C. Hutt, and S. F. Brown, “Thresholds for the Perception of Whole Body Angular Movement about a Vertical Axis,” Aviation, Space, and Environmental Medicine, vol. 60, 1989, pp. 205–213.

[19] N. Wiener, “Nonlinear Problems in Random Theory,” MIT Press, Cambridge, 1958.

[20] J. M. Maciejowski, “Predictive Control with Constraints,” Pearson Education, 2002.

[21] O. Föllinger, “Regelungstechnik: Einführung in die Methoden und ihre Anwendung (Control Engineering: Introduction to the Methods and their Application),” VDE Verlag, 2013.

[22] M. Katliar, K. N. De Winkel, J. Venrooij, P. Pretto, and H. H. Bülthoff, “Impact of MPC Prediction Horizon on Motion Cueing Fidelity,” Driving Simulation Conference Europe (DSC), 2015, pp. 219–222.

[23] M. Grottoli, D. Cleij, P. Pretto, Y. Lemmens, R. Happee, and H. H. Bülthoff, “Objective evaluation of prediction strategies for optimization-based motion cueing,” Simulation: Transactions of the Society for Modeling and Simulation International, vol. 95, no. 8, 2018, pp. 707–724.

[24] P. Biemelt, C. Link, S. Gausemeier, and A. Trächtler, “A Model-Based Online Reference Prediction Strategy for Model Predictive Motion Cueing Algorithms,” 21st IFAC World Congress, 2020.

[25] K. Svanberg, “A Class of Globally Convergent Optimization Methods Based on Conservative Convex Separable Approximations,” SIAM Journal on Optimization, vol. 12, no. 2, 2002, pp. 555–573.

[26] S. G. Johnson, “The NLopt nonlinear-optimization package,” https://nlopt.readthedocs.io/en/latest/, last call December 7th 2020.

[27] D. G. Luenberger, “Observing the State of a Linear System,” IEEE Transactions on Military Electronics, vol. 8, no. 2, 1964, pp. 74–80.

[28] N. A. Pouliot, C. M. Gosselin, and M. A. Nahon, “Motion Simulation Capabilities of Three-Degree-of-Freedom Flight Simulators,” Journal of Aircraft, vol. 35, no. 1, 1998, pp. 9–17.

[29] I. Al Qaisi and A. Trächtler, “Human in the Loop: Optimal Control of Driving Simulators and New Motion Quality Criterion,” IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012, pp. 2235–2240.

[30] P. Grant, M. Blommer, B. Artz, and J. Greenberg, “Analysing Classes of Motion Drive Algorithms Based on Paired Comparison Techniques,” Vehicle System Dynamics, vol. 47, no. 9, 2009, pp. 1075–1093.

[31] A. Siegel and H. N. Sapru, Essential Neuroscience. Lippincott Williams & Wilkins, 2006.

[32] L. R. Young and J. L. Meiry, “A Revised Dynamic Otolith Model,” Aerospace Medicine, vol. 39, no. 6, 1968, pp. 606–608.

[33] C. Fernandez and J. M. Goldberg, “Physiology of Peripheral Neurons Innervating Semicircular Canals of the Squirrel Monkey. II. Response to Sinusoidal Stimulation and Dynamics of Peripheral Vestibular System.” Journal of Neurophysiology, vol. 34, no. 4, 1971, pp. 661–675.

[34] R. J. Telban and F. M. Cardullo, “Motion Cueing Algorithm Development: Human-Centered Linear and Nonlinear Approaches,” National Aeronautics and Space Administration (NASA), CR-2005-213747, Tech. Rep., 2005.

[35] A. Nesti, C. Masone, M. Barnett-Cowan, P. Robuffo Giordano, H. H. Bülthoff, and P. Pretto, “Roll rate thresholds and perceived realism in driving simulation,” Driving Simulation Conference Europe (DSC), 2012, pp. 23–31.

Visual Customer Interaction through Emotion Detection and Face Landmarks

Rui P. Duarte*, Carlos A. Cunha*, Valter Borges*, André Ferreira* and David Mota†
*School of Management and Technology
Polytechnic Institute of Viseu, Viseu, Portugal
pduarte@estgv.ipv.pt, cacunha@estgv.ipv.pt, estgv16626@alunos.estgv.ipv.pt, af_af_10@hotmail.com
†Bizdirect Competence Center, Viseu, Portugal
david.mota@bizdirect.pt

Abstract—**Understanding consumer behavior is a dynamic field, critically important to the success of companies and to consumer satisfaction. It is especially important in scenarios of intense competition, currently characteristic of the retail store industry, where companies fight for every individual customer. A great in-store experience encourages shoppers to become loyal customers, positive word of mouth and referrals. However, the opposite happens if customers' needs are not met, a poor customer experience is provided and further visits of the customer may be at risk. Due to the dimension of several retail stores, a common problem is the location of products and the ability of customers to find them. When this occurs, sales decrease and customer satisfaction is not guaranteed, thus contributing to a poor customer experience. In this paper we present a method that targets user satisfaction, by providing the retail store a tool that detects if a product is not being found. Our approach is twofold: first, we detect if the customer is revealing signs of negative emotions by tracking the facial expressions, and second, the facial position of the customer is tracked to detect if he/she is repeatedly looking at the same place. In each context, a lost factor is updated and when a threshold is passed, the retail store assistant is notified for customer assistance. Results show that this method is well suited for emotion detection and will increase customer satisfaction and retail stores income.**

Keywords–Image recognition; sentiment analysis; activity recognition; face landmarks; user satisfaction; retail environments.

I. Introduction

Today, the development of technology has a significant impact on society and on the organizations within it. This poses significant challenges for organizations once they are obliged to keep up with developments so fast that they can often suffer if they lack the manageability. New technologies partly determine the way people relate to, and inspire the characterizations of our society. They are the new transmission channels that shape this new world, virtual and technological. Advances in technology allow organizations to be more flexible and open to change, making the most of the opportunities that appear in the market. These opportunities are partially defined by consumers, which are an increasingly visual society. Everything we see as colors, textures, shapes, and images can communicate something to us and the ability to use this type of information is of most importance for companies. In this paper, we focus on the consumer experience by providing an employee assistance to avoid consumer unsatisfactory experiences. It improves the work presented in CENTRIC'2019 [1] by adding a new method based on face landmkarks, which, coupled with the emotion detection method, increases customer satisfaction in a retail store environment.

The concept of shopping has been changing during the years [2]. Today shops are not only the place where customers go to buy products but also the place where they spend part of their time. Thereof, retail stores need to adapt to the needs of customers in order to provide them a positive experience. Two perspectives are present: the customer that wants to find and buy a specific product and the retail store that wants to increase sales. Although in real context scenarios an easy match can be established between perspectives, they have different approaches to achieve a win-win-win solution for the customer–retailer–manufacturer relation. According to Oliver, R.L. [3], it is more challenging to fidelize an existing customer than to attract new ones. However, sometimes, this is not the case: a customer enters the retail store to buy a product, does not find it, and leaves the shop without spending money on that product. This transforms the process into an unsatisfactory experience for all the players involved.

The application of Video Analytics Technology (VAT) in retail dates back more than two decades [4]. More recently, due to advances in computer vision, machine learning, and data analysis, retail video analytics can provide retailers with much more insightful business intelligence [5][6][7]. Thus it promises much higher business value, far beyond the traditional domain of security, authentication, and loss prevention. Examples of this include analysis of store traffic, queue data, customer behavior, and purchase decision making among others. However, it is a complex real-world scenario, and many technical challenges are present for realistic computer vision techniques: changing and uncontrollable lighting conditions, high-level, complex human and crowd activities, cluttered backgrounds, crowded scenes, occlusion, odd viewing angles, low resolution cameras, limited contrast, and low object discriminability [5]. It is well known that VAT mostly focuses on automatic customer detection for the retail store industry. However, customer perspective is of most importance since they acquire products available in stores. One of the potential areas of interest is to determine whether a customer is not finding a specific product. As a consequence, the customer leaves the store without buying it, which does not relate to a win-win-win situation. Thus, it is of most importance to collect more information about customers by using VAT to detect if they are not finding a product and generate triggers to employees informing of the problem. This will increase customer satisfaction and retail store sales.

This paper deals with the role of visual interaction with customers as a strategic resource to promote competitiveness and interactivity between organizations and customers. We target a library assistant, whose main objective is to capture the customer using webcam in real-time and to detect, over a time window, that they need assistance from an employee. For this, a camera has to be in continuous capture of customer facial information and two algorithms detect if a particular customer is lost and needs assistance. In addition to the main goal described above, other objectives have been identified. The specific contributions of this paper are:

- *Emotion analysis.* To our best knowledge, this is the first scalable attempt to measure negative emotions to determine if a customer is not finding products in a retail stores. These negative emotions are the basis of unsatisfactory behavior of the customer in the context of a purchase.
- *Exploration of face landmarks for customer detection.* Face landmarks feature physical characteristics of customers. From our knowledge this is the first attempt to used them to detect if a customer is lost when looking at products in a shelf. We track the position where a customer is looking at, and determine if the location is repeated.
- *Real-time notification and intervention.* An integrated web platform is developed for the real-time notification of retail stores assistants and intervention with customers when emotions are negative or repeated places where visualy repeated.

The remainder of this paper is organized as follows. Section II briefly reviews works in the field of video analytics technology. Section III details our approach to the underlying problem, and presents a two level method based on negative emotion analysis and face landmarks. Section IV presents a web interface for the retail store assistant, where methods are validated with experimental results carried out in real context scenarios. Section V concludes the paper, providing some hints to future work.

II. Related Work

Automatic detection of human emotions is a complex problem that has been applied to several ordinary problems. Techniques addressing this problem spans several types of data sources. Faces' images are one of the most promising sources for data analytics related to the emotion detection problem and to the physical behaviour of customers.

Our work overlaps with previous research on automatic analysis of human behavior inside retail stores. In this context, several approaches have been studied, like hot zone analysis, automatic activity recognition and sentimental analysis.

A. Hot Zone Analysis

Hot zone analysis aims to identify the trajectory of customers within a store. Trajectory analysis unveils spots with more activity and reveal where customers spend their time. Human's head position estimation was explored to create the initial estimates for tracking algorithms. Zhao et al. [8] presented a method for the detection and tracking of several humans in video frames. They propose boundary and shape analysis for human detection. On top of that, a 3D walking model predicts motion templates from the captured frames to track humans. This work was later improved by Zao and Ram [9], through the inclusion of a detection technique for human identification using Markov chain Monte Carlo methods. The method was tested in indoor and outdoor high-density scenes. In the outdoor scenes, false positives appear at far ends and dense edges. In the indoor scenes, the subtraction method gives erroneous foreground blobs. For human segmentation in both scenes, 1000 iterations are necessary to segment human objects. Leykinv and Mihran [10] developed a method where the human head coordinates are extracted from video frames to determine the position of customers in a store. These coordinates are further used to track customers in video sequences captured in crowded environments. The low-level extraction of the customers in a frame and the use of camera calibration to locate customer's head and location in the picture allows them to infer their location in the store.

B. Activity Recognition

The activity recognition is related to the shop behaviour and represents the actions of customers when buying products. Monitoring this behaviour is of most importance to academics and retail stores. Popa et al. [6] analyzed customer behaviour using background subtraction form images. This approach allowed them to detect customers in the entry point and then track them in the system. In [7], Popa et al. improved the method for automatic assessment of customer' appreciation of products. First, they classified customer behaviour by participant observation. Next, they implemented a model for motion detection, trajectory analysis, and face location and tracking for different customers. Sicre and Nicolas [11] resorted to behaviour models for detection of motion, tracking moving objects, and describing local motion. Results have shown that the approach can correctly classify 73% of the frames, for sequences taken in real environments. Later, Frontoni et al. [12] proposed a method to analyze human behavior in shops in order to increase consumer satisfaction and purchases. In their method, they use vertical red, green and blue depth sensors for people counting and shelf interaction analysis. Their results exhibited areas with both positive and negative interactions with products in shelves. They compared their results with ground truth visually recorded, and accuracy varies between 97.2% and 98.5%. Hu et al. [13] investigated the detection of semantic human actions in complex scenes. Their work deals with spatial-temporal ambiguities in frames using bag of instances representing the candidate regions of individual actions. A technique based on the combination of Simulated Annealing and Support Vector Machines has shown better results than standard Support Vector Machines.

C. Sentiment Analysis in Videos

Sentiment analysis is another area of video analytics. This type of problem is related to the problem addressed in this paper, since it acquires the emotional level of the customer. Zadeh et al. [14] addressed this problem using a multimodal dictionary that exploits jointly words and gestures. The approach has shown better results than straightforward visual and verbal analysis. An alternative approach to methods that adopt bag of words representations and average facial expression intensities is presented by Chen et al. [15]. They propose sentiment prediction using a time-dependent recurrent approach that performs fusion of several modalities (e.g., verbal, acoustic and visual) at every time-step. The implementation of the approach using long short-term memory networks has shown significant improvements over several other approaches. Wang and Li [16] explored sentiment analysis in social media images. The main challenge of the work lies in the semantic gap between visual features and underlying sentiments. Contextual information is proposed to overcome the semantic gap in prediction of image sentiments. The solution was shown effective when evaluated with two large-scale datasets.

III. Approach

The approach presented in this paper is based on a machine learning system that runs in background for the intervention of retail store assistants with costumers. It focuses on the analysis of information obtained from a facial recognition system at two levels: emotion analysis and face landmarks. At the emotional level, when negative emotions are detected, the retail store assistant is notified for customer intervention. At the face landmark level, when a costumer is detected to be looking at the same place several times, the intervention is triggered.

A. Problem Statement

The study of human behavior in retail stores has been carried out in the last years, and it can be interpreted by analyzing the human emotional responses to contexts [17]. Moreover, the tracking of the position of the customer face can also be used to detect patterns of customer behaviour in retail stores.

Figure 1 presents a general view of the specification of the problem at the emotional and physical levels. The example assumes that a costumer is buying a book in a bookstore and is trying to find it in a shelf. A typical behaviour consists on eye motion between books and validation if the book cover is the one that he/she is looking for. This can be represented by emotions that can be positive or negative representing the customer state of mind when looking for a product (represented by the sequential arrows). Typical physical behaviour is also related to the repetition of a position in situations that the product is not being found (the gray circle represents the moment a customer looks more than once to the same place). If these one of these two characteristics are detected, it implies that a costumer is not finding a product and a retail store assistant can go to the customer for assistant.

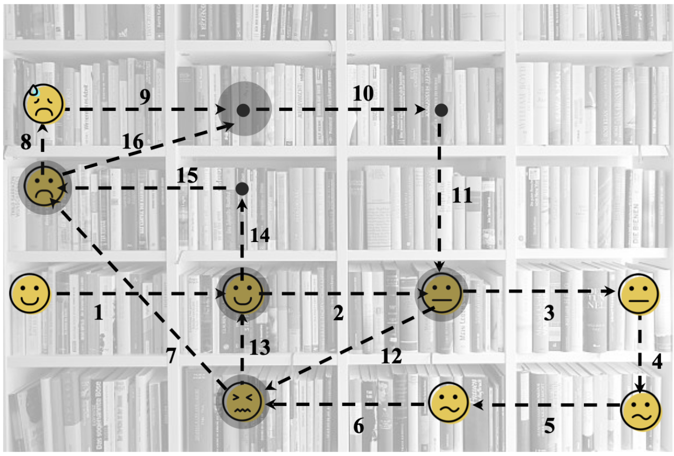


Figure 1. Problem specification for emotion and physical analysis.

At the emotional level, one of the problems that currently exist in customer service is trying to understand their state of mind when inside a store. For that purpose, the detection of emotions from customers will be able to increase the quality of service - the more relevant information about the customer, the better the assistance. The measurement of emotions can be carried out by several applications that are available in the market. These emotions can be either negative or positive. This work aims at the detection of negative emotions in a time window, where sadness is one of the most significant negative emotion to consider. However, manifestation of negative emotions can also be measured using other parameters like anger, disgust, or fear. In this paper we explore the combination of several negative emotions to determine a sadness level, β, used for costumer intervention.

Moreover, at the physical level, this work aims at providing a tool to explore face landmarks that are detected within a repeated context of interaction. Here, when a customer is not finding a product, it is normal that he/she looks around or starts to make random movements, which are indicators of uncertainty. This type of head movement can be captured using facial recognition software and can be used to detect if a costumer is looking at the same products he/she looked before, which may indicate the need for customer intervention, by determining a recurrent level, ρ.

Thus, tracking negative emotions and physical characteristics of faces in the context of a store are open problems, which is of most importance to be solved since they serve the automation of customer-employee contexts, resulting in an increase of the speed of attendance, improve customer satisfaction and increase retail stores sales.

B. Machine Learning Implementation

The performance of machine learning models is deeply dependent on the volume of data available for training models. For that reason, the most accurate models are provided by giants of software that have access to large volumes of data for training models capable of accurate detection of emotions in images. Fortunately, these models are widely available through an Internet accessible API like the IBM Watson [18], Face API [19], Kairos [20], and Amazon Rekognition [21].

In this work, we use Face API [19]. It is a cognitive service developed by Microsoft that provides algorithms to detect, recognize, and analyze human faces in images. Face API features are obtained in two stages: the first is the detection and recognition of face attributes; in the second, a JSON file is returned with the fields that contain face attributes.

Let $\mathcal{C} = \{c_j\}$, $j = 1 \ldots M$, be the number of customers that are detected in the system and $\mathcal{F} = \{f_i\}$, $i = 1 \ldots N$, the number of frames captured in real-time using the Face API for each customer $c_j \in \mathcal{C}$. The detection stage represents the analysis of the existing faces, $\mathcal{F}$, of customers, $\mathcal{C}$, and returns attributes for each $\{f_i\}$. When $\{f_i\}$ is detected, the face rectangle attribute is returned, since it contains the pixels to track $\{f_i\}$ in the image and gets its bounding box.

Within this bounding box, other attributes are returned by the API to the JSON file, namely, face Id, face landmarks, age, emotion, gender, and hair. In this paper all the parameters

TABLE I. USER TESTING IN REAL SCENARIOS: ACTING NORMAL, SIMULATION, FORCE SADNESS, FORCE ANGER AND FORCE HAPPINESS.

Anger (A_p)	Contempt (C_p)	Disgust (D_p)	Fear (F_p)	Happiness (H_p)	Neutral (N_p)	Sadness (S_p)	Surprise (Su_p)	Testing
0	0.001	0	0	0	0.999	0	0	*acting normal*
0.001	0.001	0	0	0	0.985	0.014	0	*simulate scenario*
0	0.002	0	0	0	0.762	0.235	0	*force sadness*
0.004	0.005	0.005	0	0.001	0.962	0.022	0	*sumulate scenario*
0.005	0.002	0.001	0	0.001	0.731	0.261	0	*force sadness*
0	0.002	0	0	0	0.993	0.005	0	*acting normal*
0	0	0	0	1	0	0	0	*force happiness*
0	0.016	0	0	0	0.811	0.172	0	*force sadness*
0.031	0.001	0	0	0	0.967	0.001	0	*simulate scenario*
0.035	0.001	0	0	0	0.966	0.001	0	*force anger*
0	0	0	0	0	0.977	0.023	0	*simulate scenario*
0	0.001	0	0	0	0.905	0.094	0	*simulate scenario*
0	0	0	0	0	0.958	0.041	0	*force sadness*
0	0.089	0.001	0	0	0.58	0.33	0	*force sadness*
0.001	0.027	0	0	0	0.967	0.004	0	*acting normal*
0	0.152	0	0	0.848	0	0	0	*force happiness*
0.172	0.002	0	0	0	0.823	0.003	0	*force sadness*
0.011	0.006	0	0	0	0.962	0.021	0	*simulate scenario*
0.008	0.37	0	0	0	0.621	0.001	0	*force anger*
0.16	0.043	0.001	0	0.001	0.661	0.134	0	*simulate scenario*
0.001	0.025	0	0	0	0.967	0.007	0	*simulate scenario*
0	0.169	0	0	0.009	0.821	0	0	*force sadness*
0.0058	0.011	0	0	0	0.887	0.043	0	*force sadness*
0	0.004	0	0	0.006	0.987	0.004	0	*acting normal*
0	0.001	0	0	0.958	0.04	0.002	0	*force happiness*
0	0	0	0	0	0.857	0.143	0	*force sadness*
0	0	0	0	0	0.84	0.159	0	*simulate scenario*
0.412	0.042	0.09	0.029	0.006	0.57	0.001	0.363	*force anger*
0.001	0.007	0	0	0.001	0.94	0.051	0	*simulate scenario*
0	0.005	0	0	0.038	0.955	0.001	0	*simulate scenario*
0	0.001	0	0	0	0.958	0.041	0	*force sadness*
0	0.001	0	0	0	0.417	0.582	0	*force sadness*
0	0	0	0	0	0.997	0.002	0	*acting normal*
0	0	0	0	1	0	0	0	*force happiness*
0	0	0	0	0	0.965	0.035	0	*force sadness*
0.053	0.004	0	0	0	0.943	0	0	*simulate scenario*
0.127	0.009	0	0	0	0.864	0	0	*force anger*
0	0.0087	0	0	0.036	0.868	0.002	0.007	*simulate scenario*
0	0.001	0	0.001	0.001	0.956	0.033	0.009	*simulate scenario*
0	0.003	0	0	0	0.679	0.318	0	*force sadness*
0	0	0	0	0	0.887	0.113	0	*force sadness*

are considered in three contexts. First, for a general characterization of the costumer, age, gender, and hair attributes are used. These attributes allow the retail store employee to better identify the customer (note that for security policies, the system cannot store the face of the customer). Next, for the emotion analysis (cf., Section III-C), the emotion attribute, containing a set of different emotions, is used to detect negative emotions. Finally, at the facial level (cf., Section III-D), face landmarks are used to track repetition of previously visited positions.

The parameters returned by the Face API are a basis of knowledge for the implementation of the emotion and facial tracking methods presented in the following sections.

C. Emotion Analysis

There are several parameters associated to emotions that are returned by facial recognition systems, namely anger (A_p), contempt (C_p), disgust (D_p), fear (F_p), happiness (H_p), neutral (N_p), sadness (S_p) and surprise (Su_p). In the scope of this work, we only consider negative emotions (A_p, D_p, F_p and S_p) that affect the costumer interaction with the system.

The basic idea of our method is presented in Figure 1 (which includes both representations of emotions and physical motion). When a customer arrives at a shelf, Face API captures his emotions, and a sadness level β is set to zero. This factor updates in the presence of negative emotions, and once a threshold is passed ($\beta > 50\%$), the assistant is asked to go to the customer. Negative emotions manifest in several ways, and one of the most critical parameters is the sadness parameter, $S_p \in [0..1]$ (values near 1 correspond to the total manifestation of sadness). Therefore, once a frame captures a customer with a high value of sadness, it may be an indicator of a potential product not being found by a customer. Other parameters like A_p, D_p or F_p are also present in negative emotions, and their contribution is analyzed in this paper.

To determine the weights to consider in each of the negative emotions, an empiric study (presented in Table I) was carried out with users that were asked to express several emotions: S_p, N_p, D_p, H_p and simulate the action of looking for a product and not finding it, referred to as *Simulated*. In the emotion tests considering H_p and N_p, these parameters have high values, representative of the tested emotion. In the tests for forced sadness and simulation, S_p has low values in most cases, which is justified by the fact that the sadness emotion can result in false positives. However, in this case, the presence of other negative emotions is visible, with small values of A_p, D_p and F_p. Analyzing the impact of these parameters in the emotion is an essential factor to determine how to infer sadness when S_p should be naturally present and is not.

In this context, two types of tests were carried out: first, the evaluation of the impact of each negative emotion and, second, the presence of all negative emotions. In the first test, results obtained ($A_p = 47\%$, $D_p = 16\%$, $F_p = 6\%$ and $S_p = 91\%$), show that negative emotion is present in the tests. However, excluding S_p, the other negative emotions are not feasible to be used individually to complement the sadness test, since they are present in a small number of tests, which are not representative of the sample. In the second test, was considered the cumulative presence of all negative emotion parameters $(A_p + D_p + F_p + S_p > tol)$ for the same scenario (forced sadness and simulation), as shown in Table II.

TABLE II. TOLERANCE TESTS FOR $A_p + D_p + F_p + S_p > tol$

	Tolerance (tol)				
	0.0	0.01	0.02	0.03	0.04
Cumulative negative emotions (%)	97.22	83.73	80.16	74.32	68.26

Results show that when $tol = 0.0$, 97.22% of the tests reveal the presence of cumulative negative emotions, which is very representative of the tested scenario. The rate decreases for $tol \geq 0.01$. Therefore, when S_p is not representative in a sadness test, the alternative of considering cumulative negative emotions has success rate of 97.22%. Recall that these criteria are used only to improve the success rate of retail store assistants interventions and are used in two contexts: in the evident presence of sadness (high values of S_p) and in the presence of signs of sadness ($A_p + D_p + F_p + S_p > tol$, for low values of S_p). The resulting method is presented in Algorithm 1.

Algorithm 1: Emotion-based intervention method

```
Data: $\mathcal{C}$                    ◁ detected customers $\mathcal{C} = \{c_j\}$
Data: $\mathcal{F}$                              ◁ API frames $\mathcal{F} = \{f_i\}$
Result: $\beta$, $\mathcal{I}$      ◁ $\beta$ = sadness level, $\mathcal{I}$ = Intervention
begin
    foreach $c_j \in \mathcal{C}$ do
        $\beta_j \leftarrow 0.0$                  ◁ set sadness level to zero
        $\mathcal{I} \leftarrow false$              ◁ no intervention required
        foreach $f_i \in \mathcal{F}$ do
            $A_{p_i} \leftarrow A_p \in f_i$          ◁ get anger from $f_i$
            $F_{p_i} \leftarrow F_p \in f_i$            ◁ get fear from $f_i$
            $S_{p_i} \leftarrow S_p \in f_i$        ◁ get sadness from $f_i$
            $D_{p_i} \leftarrow D_p \in f_i$        ◁ get disgust from $f_i$
            if $(S_{p_i} > 0.5)$ then
                $\beta_j \leftarrow \beta + 0.1$      ◁ update sadness level
            else if $(A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0)$ then
                $\beta_j \leftarrow \beta + 0.05$   ◁ update sadness level
            if $(\beta_j > 0.5)$ then
                $\mathcal{I} \leftarrow true$         ◁ intervention required
        end
    end
end
```

The algorithm starts by scanning if a customer is detected by the Face API and its faceId is generated. The sadness level of each customer, β_j, is set to zero, and frames are captured while the customer is detected in the system. For every captured frame, the Face API returns negative emotion values that are stored for processing. Every time the algorithm captures evidence of sadness ($S_{p_i} > 0.5$ or signs of sadness ($A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0$), the value of β_j is updated in a factor of 0.1 or 0.05, respectively. When the sadness level passes a threshold of 0.5, the assistant is informed that a customer needs intervention.

An important consideration is that our system does not retain personal information of a customer. After detection by a camera, only a faceId is generated to uniquely identify the characteristics of that customer. If he/she leaves the system, the method still continues to try to track the faceId of the customer for five minutes. After that period, the information of the faceId is removed from the database, but the face attributes are kept. With this, personal information of users is not stored, therefore, it does not allow the system to track a specific customer. If the customer is again detected in the system, he/she will be assigned a new faceId.

D. Physical Motion Analysis

At the physical level, a human face is composed of sets of points that can be well identified. These points, called face landmarks go from pupils to the tip of the nose. Face landmark detection is a computer vision technique developed automatically detect some particular landmarks in human faces using machine learning algorithms. The accurate identification of facial landmarks is a process by which a number of complicated image analysis problems are solved. This identification has been extended outside the domain of image research and into other applications, such as the medical field [22][23], animation [24][25], face reconstruction [26] and security [27]. In [28] and [29], a complete review of facial landmark identification techniques is presented.

Face API features 27 predefined landmark points in a face describing physical characteristics of a face: eyebrows, eyes, nose and mouth. In this paper we explore the landmarks associated to the nose, more concretely, the nose tip. In the Face API, the nose tip is captured in (x, y) coordinates, which is important to detect motion in a captured frame. In this context, the method presented in Algorithm 2 detects if a customer is looking at a (x^*, y^*) point in the neighborhood of a previous point (x, y) captured in a previous frame (see Figure 1). If that occurs, the likelihood of a product not being found increases.

As in Section III-C, once a costumer is detected, a faceId is generated, and the recurrent level, ρ, is set to zero. As a customer looks for products, new frames are captured and nose tip coordinates are determined. In this context, let f_j be the present frame and (x^*_j, y^*_j) the the nose tip corresponding coordinates. For each capture frame, if it is in the neighbourhood of a previous face ($|x^*_a - x_p| < tol$ and $|y^*_a - y_p| < tol$), then it is assumed that the customer is looking at the same place. This increases ρ in 0.01 until $\rho > 0.5$, and the retail store assistant receives notification for intervention.

Algorithm 2: Physical-based intervention method

```
Data: $\mathcal{C}$                         ◁ detected customers $\mathcal{C} = \{c_j\}$
Data: $\mathcal{F}$                               ◁ API frames $\mathcal{F} = \{f_i\}$
Result: $\rho$, $\mathcal{I}$      ◁ $\rho$ = recurrent level, $\mathcal{I}$ = Intervention
begin
    foreach $c_j \in \mathcal{C}$ do
        $\rho_j \leftarrow 0.0$                   ◁ set recurrent level to zero
        $\mathcal{I} \leftarrow false$                ◁ no intervention required
        $f_j \leftarrow$ Get Current Frame
        $N_{x^*} \leftarrow N_x \in f_j$      ◁ get nose tip $x$ coordinate
          from $f_j$
        $N_{y^*} \leftarrow N_y \in f_j$      ◁ get nose tip $y$ coordinate
          from $f_j$
        foreach $f_i \in \mathcal{F}$ do
            $N_x \leftarrow N_y \in f_i$  ◁ get nose tip $x$ coordinate
              from $f_i$
            $N_y \leftarrow N_y \in f_i$  ◁ get nose tip $y$ coordinate
              from $f_i$
            if $(|N_{x^*} - N_x| < tol$ and
            $|N_{y^*} - N_y| < tol)$ then
                $\rho_j \leftarrow \rho_j + 0.1$  ◁ update recurrent level
            if $(\rho_j > 0.5)$ then
                $\mathcal{I} \leftarrow true$          ◁ intervention required
        end
    end
end
```

IV. Experimental Design and Results

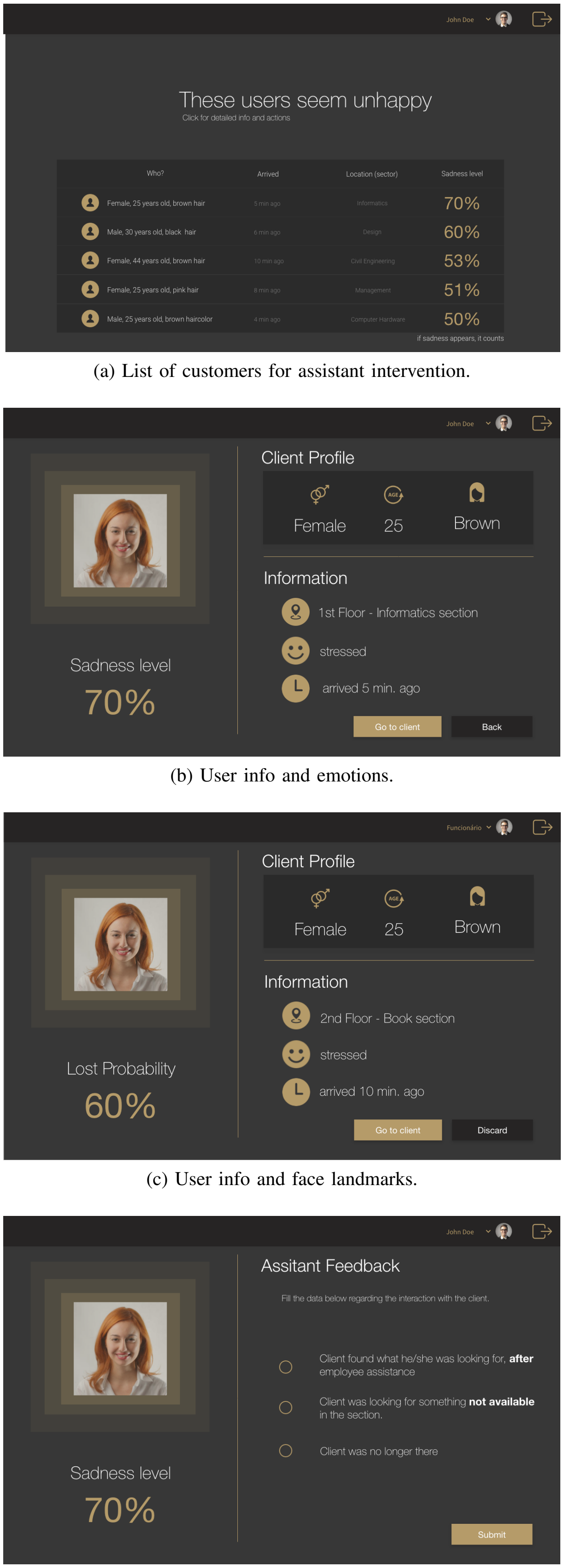
The algorithm presented in the previous section runs in background and processes information that can be visualized by the retail store assistant in an web application (see Figure 2).

A. Web Interface

The design and implementation of a web interface for the retail store assistant was of must importance to carry out a pilot study. The assistant has access to the notifications management page, presented in Figure 2a). This page is updated in real time and contains a list of customers that require intervention. Here, some general information of the customers is provided for better identification.

When the assistant selects a customer for intervention, Algorithm 1 and Algorithm 2 (running in background) stop increasing β and ρ for that customer, respectively, and these values are stored. Otherwise, they would reach the threshold value for all customers in the time elapsed between the interaction and the time to go to the customers. When the assistant selects a customer, general information is provided (such as hair color, age, gender, location in store, the emotion revealed by the customer and how long the customer is in the system). In addition, the assistant has the possibility to attend the customer or to cancel and return to the call management page as shown in Figure 2b) and Figure 2c).

The intervention level starts when the assistant clicks in the ”go to client” button and the page changes so that feedback data can be provided by the assistant, which possesses relevant information regarding the intervention with the customer, as shown in Figure 2d). It is important to note that while the



These users seem unhappy

Click for detailed info and actions

Who?	Arrived	Location (sector)	Sadness level
Female, 25 years old, brown hair	5 min ago	Informatics	70%
Male, 30 years old, black hair	6 min ago	Design	60%
Female, 44 years old, brown hair	10 min ago	Civil Engineering	53%
Female, 25 years old, pink hair	8 min ago	Management	51%
Male, 25 years old, brown haircolor	4 min ago	Computer Hardware	50%

if sadness appears, it counts

(a) List of customers for assistant intervention.

Client Profile

Female 25 Brown

Information

1st Floor - Informatics section

stressed

arrived 5 min. ago

Sadness level

70%

(b) User info and emotions.

Client Profile

Female 25 Brown

Information

2nd Floor - Book section

stressed

arrived 10 min. ago

Lost Probability

60%

(c) User info and face landmarks.

Assitant Feedback

Fill the data below regarding the interaction with the client.

○ Client found what he/she was looking for, **after** employee assistance

○ Client was looking for something **not available** in the section.

○ Client was no longer there

Sadness level

70%

(d) Assistant feedback.

Figure 2. Web interface for retail store assistant intervention.

assistant is attending the customer, no further changes in the customer emotions are captured. It is intended to capture the emotions that have caused the customer to exceed the emotional threshold and not to register emotion changes while being under intervention.

B. Results

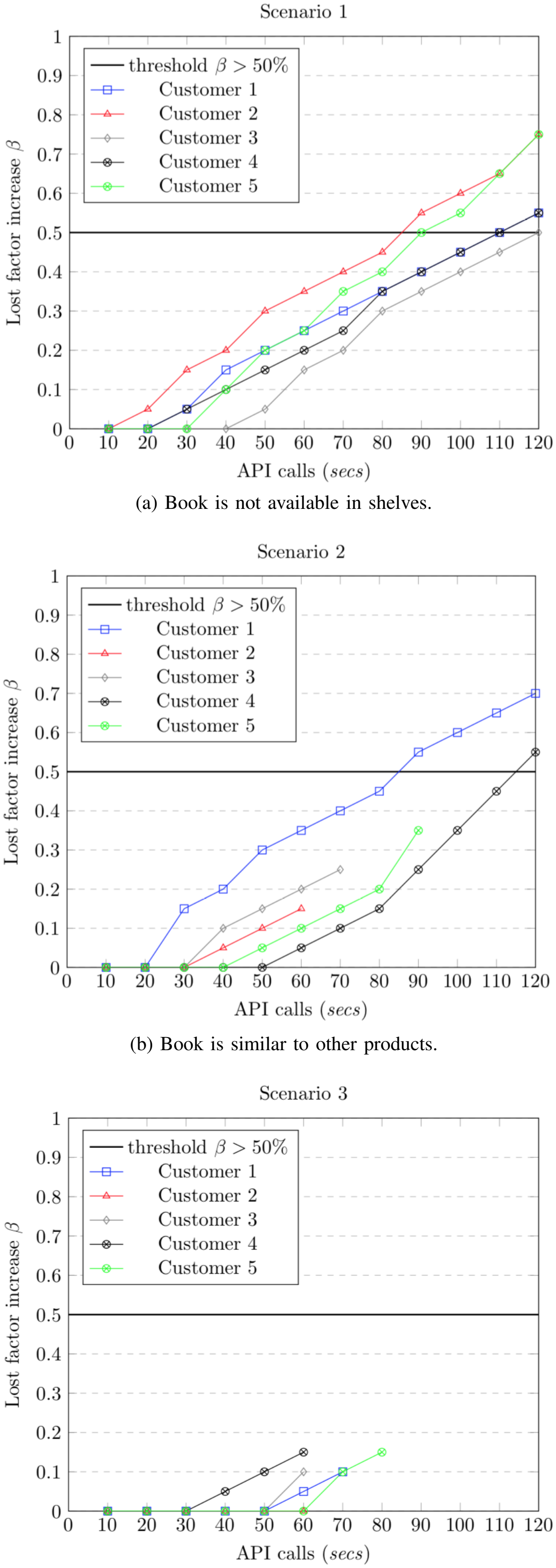
We have tested the approach described in Section III by carrying out a pilot study at both the emotional and physical levels. Books were placed in shelves with a camera placed to capture emotions and face landmarks. Five customers where asked to find a book, from twenty available books, in three scenarios:

- Scenario 1: The book is not available in the products placed in the shelves.
- Scenario 2: The book is in the shelves, but very similar to other books, making it difficult to be found.
- Scenario 3: The book is available in the shelves and easy to be identified.

Results obtained are presented in Figure 3 and Figure 4, which refer to the emotion detection and face landmark detection, respectively. To provide flexibility to the system, the assistant can decide the moment of the intervention. As previously referred, when the sadness level threshold is passed, the assistant web page is updated with the customer information. However, if the assistant considers that the sadness level is not increasing with time, he/she can decide not to go to the customer. However, if the customer continues to reveal cumulative negative emotions or head motion is present, the assistant then makes the decision to assist him. Moreover, if all assistants are occupied, the system continues to increase the lost levels of a customer, until an assistant is available.

At the emotional level, for scenario 1 (Figure 3a)), costumers reveal signs of cumulative unhappiness, $(A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0)$, or sadness $(S_p > 50\%)$ as they realize that they are not finding the product. The sadness level threshold is passed for all customers after a few iterations of API calls. The variation of the sadness level cumulative response is due to the fact that, in the API calls, the customer can reveal one of both negative emotions tested. This implies that there can be an increase of 0.05 or 0.1, depending on the most prevalent negative emotion in each detection. In this context, the web interface for the assistant is updated with the data related to the new customer that requires intervention (see Figure 3a)). For all the customers of the tested scenarios, the assistant reported option two in the feedback page (see Figure 2d)). In scenario 2, three customers found the product, after some iterations and left the system. The other two reached the sadness level threshold. For them, the assistant reported option one in the feedback page. Finally, in scenario 3, all the customers found the product after a few iterations of API calls, never reaching the sadness level threshold, thus, not requiring assistant intervention.

At the physical level, the same scenarios were considered and different customers were asked to carry out the study. Figure 4 presents the results obtained by applying Algorithm 2, and similar results were obtained, when compared to the emotional tests. However, more API calls were required in

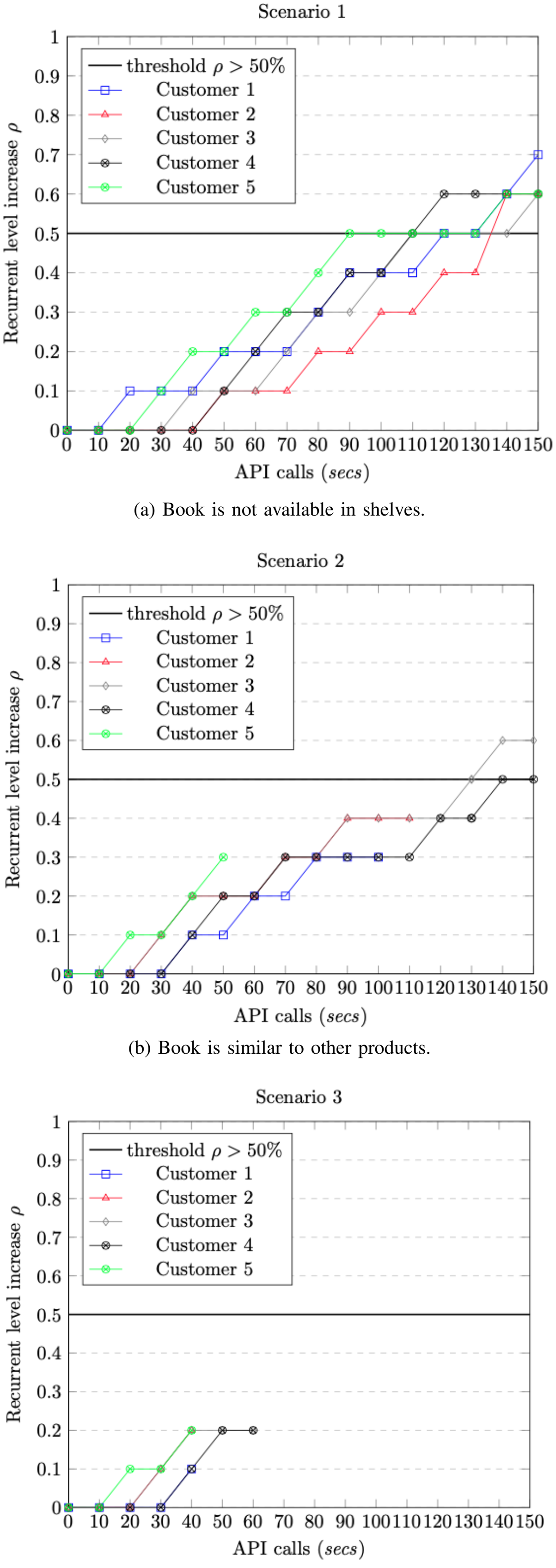


(a) Book is not available in shelves.

(b) Book is similar to other products.

(c) Book is well identified in shelves.

Figure 3. Results obtained for emotion tests with five customers in the three scenarios.

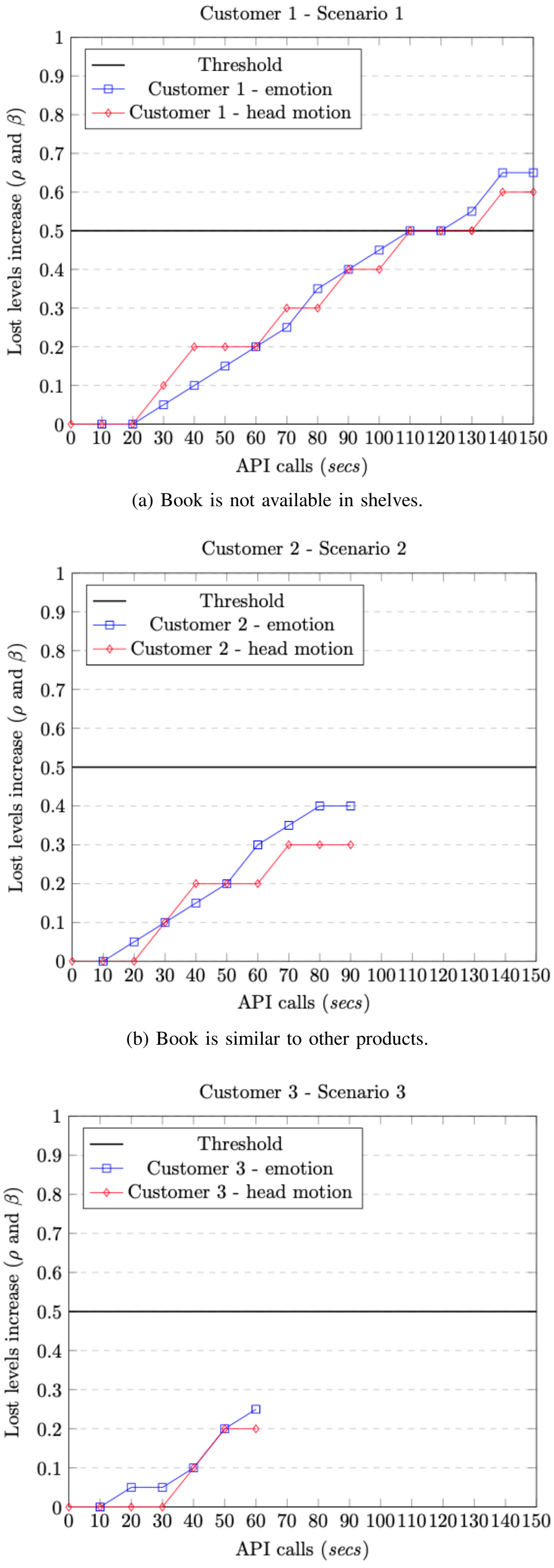


(a) Book is not available in shelves.

(b) Book is similar to other products.

(c) Book is well identified in shelves.

Figure 4. Results obtained for the detection of face position for five customers in the three scenarios.



(a) Book is not available in shelves.

(b) Book is similar to other products.

(c) Book is well identified in shelves.

Figure 5. Emotion and face analysis for customers in the three scenarios.

some tests to achieve customer intervention. In the context of Scenario 1, depicted in Figure 4a), after some iterations, customers started to look at previously visited places, which increased the recurrent level, ρ, and forced it to pass the threshold, which implied in customer intervention. In the other scenarios (Figures 4b) and 4c)), results show that most customers found the book. Results show that the method detects if a customer is not finding a product.

A final test was carried out with three customers to compare the accuracy of each method. Each customer was asked to find a book in the context of the defined scenarios. Results are presented in Figure 5 and reveal a correlation between emotion analysis and the tracking of previously visited places, for all the tested scenarios.

V. Conclusion and Future Work

This paper presented two novel scalable methods based on visual recognition of customer emotions and face landmarks when buying products, using Face API. The method uses a camera to capture the manifestation of negative emotions at two levels: the effective manifestation of sadness and evidence of sadness, in a set of frames. Concurrently, the method detects if the customer is looking at previously visited places, by extracting face landmarks. The evaluation methodology shows that both methods present good results in real scenarios. Additionally, the implementation of an intuitive web interface allows retail shops assistants to carry out interventions with customers, if the emotional and recurrent thresholds are passed. This interface will greatly assist retail stores to have an understanding of which customers require intervention and provide the necessary help in real-time. The natural implications are an increase in sales and customer satisfaction.

Future work will follow two directions, mostly focused on Artificial Intelligence (AI). A first approach will use to anticipate the needs of customers based on the previous emotional analysis. This will allow retail stores to determine which products are not being found and reorganize stores in order to better allow the correct identification of products. Moreover, the lost levels (sadness and recurrent) were obtained empirically. It will be essential to use AI as a mean to adjust these parameters.

Acknowledgments

This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Research Centre in Digital Services (CISeD) and the Polytechnic of Viseu.

References

[1] V. Borges, R. P. Duarte, C. A. Cunha, and D. Mota, “Are you lost? using facial recognition to detect customer emotions in retail stores,” *CENTRIC 2019 : The Twelfth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, p. 49–54, Nov. 2019. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=centric_2019_3_30_30031

[2] M. H. Moss, *Shopping as an entertainment experience*. Lexington Books, 2007.

[3] R. L. Oliver, “Whence consumer loyalty?” *Journal of Marketing*, vol. 63, no. 4, pp. 33–44, 1999, ISSN: 00222429.

[4] R. M. Bolle, J. H. Connell, N. Haas, R. Mohan, and G. Taubin, “Veggievision: A produce recognition system,” in *Proceedings Third IEEE Workshop on Applications of Computer Vision (WACV’96)*. IEEE, Dec. 1996, pp. 244–251, ISBN: 0-8186-7620-5.

[5] J. Connell, Q. Fan, P. Gabbur, N. Haas, S. Pankanti, and H. Trinh, “Retail video analytics: an overview and survey,” in *Video Surveillance and Transportation Imaging Applications*, vol. 8663. International Society for Optics and Photonics, 2013, p. 86630X.

[6] M. Popa, L. Rothkrantz, Z. Yang, P. Wiggers, R. Braspenning, and C. Shan, “Analysis of shopping behavior based on surveillance system,” in *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, Oct. 2010, pp. 2512–2519, ISBN: 978-1-4244-6588-0.

[7] M. C. Popa, L. Rothkrantz, C. Shan, T. Gritti, and P. Wiggers, “Semantic assessment of shopping behavior using trajectories, shopping related actions, and context information,” *Pattern Recognition Letters*, vol. 34, no. 7, pp. 809–819, May 2013, ISSN: 0167-8655.

[8] T. Zhao, R. Nevatia, and F. Lv, “Segmentation and tracking of multiple humans in complex situations,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 2. IEEE, Dec. 2001, pp. II–II, ISBN: 0-7695-1272-0.

[9] T. Zhao and R. Nevatia, “Stochastic human segmentation from a static camera,” in *Workshop on Motion and Video Computing. Proceedings.* IEEE, Dec. 2002, pp. 9–14, ISBN: 0-7695-1860-5.

[10] A. Leykin and M. Tuceryan, “A vision system for automated customer tracking for marketing analysis: Low level feature extraction,” in *Human Activity Recognition and Modelling Workshop*, vol. 3. Citeseer, 2005, pp. 6–13.

[11] R. Sicre and H. Nicolas, “Human behaviour analysis and event recognition at a point of sale,” in *2010 Fourth Pacific-Rim Symposium on Image and Video Technology*. IEEE, Nov. 2010, pp. 127–132, ISBN: 978-1-4244-8890-2.

[12] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, “Customers’ activity recognition in intelligent retail environments,” in *New Trends in Image Analysis and Processing (ICIAP 2013)*. Springer Berlin Heidelberg, 2013, pp. 509–516, ISBN: 978-3-642-41190-8.

[13] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, “Action detection in complex scenes with spatial and temporal ambiguities,” in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 128–135, ISBN: 978-1-4244-4420-5.

[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, Nov. 2016, ISSN: 1941-1294.

[15] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, “Multimodal sentiment analysis with word-level fusion and reinforcement learning,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI ’17)*. New York, NY, USA: ACM, 2017, pp. 163–171, ISBN: 978-1-4503-5543-8.

[16] Y. Wang and B. Li, “Sentiment analysis for social media images,” in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1584–1591, ISBN: 978-1-4673-8493-3.

[17] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[18] “Ibm watson - visual recognition,” URL: https://www.ibm.com/watson/services/visual-recognition/ [accessed: 2020-11-10].

[19] “Microsoft cognitive services: Face api,” 2019, URL: https://azure.microsoft.com/en-us/services/cognitive-services/face/ [accessed: 2020-11-10].

[20] “Kairos apis and sdks,” 2019, URL: https://www.kairos.com/ [accessed: 2020-11-10].

[21] “Amazon rekognition - video and image,” URL: https://aws.amazon.com/rekognition [accessed: 2020-11-10].

[22] D. L. Guarin, J. Dusseldorp, T. A. Hadlock, and N. Jowett, “A machine learning approach for automated facial measurements in facial palsy,” *JAMA facial plastic surgery*, vol. 20, no. 4, pp. 335–337, 2018.

[23] A. T. Balaei, K. Sutherland, P. A. Cistulli, and P. de Chazal, “Automatic detection of obstructive sleep apnea using facial images,” in *2017 IEEE*

14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, 2017, pp. 215–218.

[24] K. Liu, A. Weissenfeld, J. Ostermann, and X. Luo, “Robust aam building for morphing in an image-based facial animation system,” in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 933–936.

[25] S. Ioannou, G. Caridakis, K. Karpouzis, and S. Kollias, “Robust feature detection for facial expression recognition,” *Journal on image and video processing*, vol. 2007, no. 2, pp. 5–5, 2007.

[26] U. Park and A. K. Jain, “3d face reconstruction from stereo video,” in *The 3rd Canadian Conference on Computer and Robot Vision (CRV’06)*. IEEE, 2006, pp. 41–41.

[27] C. Pradhan, D. Banerjee, N. Nandy, and U. Biswas, “Generating digital signature using facial landmlark detection,” in *2019 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2019, pp. 0180–0184.

[28] B. Johnston and P. de Chazal, “A review of image-based automatic facial landmark identification techniques,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 86, 2018.

[29] O. Celiktutan, S. Ulukaya, and B. Sankur, “A comparative study of face landmarking techniques,” *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 13, 2013.

Structural Equation Modeling with Sentiment Information and Hierarchical Topic Modeling

Takurou Ogawa
Department of Sustainable System Sciences
Graduate School of Humanities and Sustainable Systems
Osaka Prefecture University
Japan
e-mail: saa01052@edu.osakafu-u.ac.jp

Ryosuke Saga
Department of Sustainable System Sciences
Graduate School of Humanities and Sustainable Systems
Osaka Prefecture University
Japan
e-mail: saga@cs.osakafu-u.ac.jp

Abstract—**Service evaluation depends on various factors, such as assurance, responsiveness, and tangibles. Given that emotional satisfaction affects service satisfaction, analyzing both the evaluation and sentiments is important in improving service. Previous studies have identified the evaluation factor and determined the degree of influence on the resulting evaluation. However, there is little effective analysis that reflects the influence of such a factor on sentiment. In this study, we use hierarchal Latent Dirichlet Allocation and structural equation modeling (SEM) to express the causality relationships of service evaluation visually and quantitatively. Sentiment obtained quantitatively by using sentiment analysis is newly applied to SEM to obtain knowledge reflecting the influence of sentiment. As a result of the experiment, we can identify the causality of service and determine the influence of the evaluation factor and sentiment quantitatively. Furthermore, we conduct an experiment that compares a causal model with and without sentiment information and improve the model interpretability.**

Keywords-sentiment analysis; service analysis; structural equation modeling; hierarchical Latent Dirichlet Allocation; causal analysis

I. INTRODUCTION

In recent years, the service industry has grown rapidly such that in developed countries, there are so many markets that account for 60% to 70% of a country's gross domestic product (GDP). In the United States where GDP is the highest, the service industry's GDP is $ 15.52 trillion, accounting for 80% of the total GDP [1][2]. In addition, with the spread of smartphones, apps for various services (e.g., Twitter, navigation), the introduction of recommended hotels, and the rise of electronic services (e.g., Internet shopping) are rapidly increasing. With this background, the importance of services has grown in recent years. Service improvement is important as services are produced and consumed at the same time compared with products that are released and finished. Thus, analyzing the evaluation of the service in order to improve such service is important.

Service evaluation depends on various factors, such as assurance, responsiveness, and tangibles. For example, SERVQUAL evaluates the quality of service [3] with five-dimensional indicators, and Airport Service Quality [4] defines airport evaluation factors. As there are many factors in the evaluation of services, it is necessary to find out the evaluation factors to analyze the evaluation.

Generally, analyzing services is difficult because these have special features that ordinary products do not have like Intangible, Heterogeneous, Inseparable, and Perishable. However, there are several clues to analyze the services from the data (e.g., questionnaire). Especially, user review is useful because the review describes user experience of and perceived from the services. It is possible to analyze the quality of service and the evaluation of service. Meanwhile, emotional satisfaction is also regarded as an important and attractive factor in service satisfaction. That is, customers experience different positive and negative sentiments related to service, and these sentiments influence service satisfaction [5]. Of course, these factors influence service evaluation and the sentiments related to the service are implied in the user review; however, there is no study to identify and analyze evaluation factors together with sentiment information.

This paper describes the method by which to perform causality analysis from text data, such as user review. In order to treat causal analysis, we use the topic-based approaches by applying a topic model to the review. In addition, the sentiments for evaluation factors in the text are quantitatively determined using sentiment analysis method to understand emotional satisfaction. By applying topic and sentiment information to structural equation modeling (SEM), we analyze the influence of each factor quantitatively.

The first contribution of this paper is that it obtains the knowledge reflecting sentiment information from the user review by using sentiment analysis. Second, it understands the influence on the sentiment of the evaluation factor based on the idea that sentiments are essential for service evaluation factor analysis. By using SEM with path diagram, we can also analyze and understand the causality relationships among topics and their sentiments associated with topics that are visually and quantitatively express.

This paper is structured as follows. Section II refers to the existing related research, Section III explains the core method of the analysis process, and Section IV describes analysis experiments using actual data. Finally, Section V discusses future work and Section VI concludes this study.

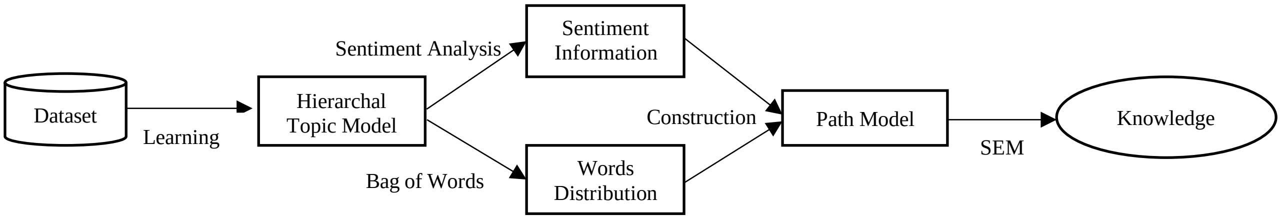


Figure 1. Analysis process

II. Literature Review

In related research for service analysis, SERVQUAL [3] measures the quality of service by measuring the gap between advance expectation and subsequent experience using five indicators prepared in advance. SURVPERF [6] measures the quality of service based on the subsequent experience alone. Related researches include a study that further increased the dimension from these five dimensions [7] and another that changed the dimension to measure the quality of electronic service [8]. There are many evaluation indicators, but it is difficult to measure all services by one standard because there are many types of services and their characteristics largely differ.

Meanwhile, related works on SEM include a study that has found relationships between customer loyalty and service quality [9] and another that has proposed a model to infer the purchase factor of the game by combining hierarchal Latent Dirichlet Allocation (hLDA) and SEM [10]. A previous work used SERVQUAL and SEM to examine the effects of the former [11]. A study increased dimensions of the SERVQUAL and analyzed it through SEM [7]. Another study identified the factors that affect customer satisfaction and the dimensions of service quality and their ranking in the context of fast food restaurants [12]. A previous study used the main aspects of pedestrian level of service (PLOS) [13], namely, safety, security, mobility and infrastructure, and comfort and convenience, to provide a comfortable and safe walking environment. PLOS is a measurement tool for evaluating the degree of pedestrian accommodation on roadways. This study also used SEM to provide the essential information for interpreting the aspects of the walking environment that influence PLOS [14]. Another research analyzed the influence of e-commerce services, which are the core dimension of e-service quality, on internet banking adoption and brand loyalty of banks [15]. These works, however, do not consider the sentiment contained in the text.

Meanwhile, emotional satisfaction is largely believed to affect service satisfaction [5]. In relation to this, sentimental analysis is useful in comprehending and handling the sentiment information. A study utilizes sentiment analysis and Latent Dirichlet Allocation (LDA) to evaluate the quality of airport services [16], while another determines the user's evaluation for each attribute by combining Airport Council International-defined airport service quality attributes and sentiment analysis [17]. In these studies, sentiment is considered one of the important factors in sales of services; thus it is essential to consider sentiment. However, no study has proposed structural equation modeling that considers the sentiment contained in text.

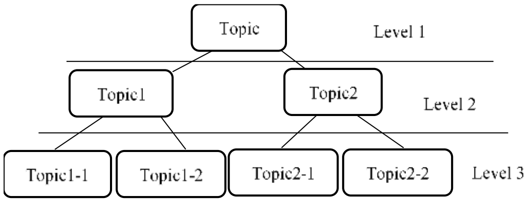


Figure 2. Hierarchal structure of topics

Therefore, the current paper proposes the model for SEM with sentiment information. By using this model, we can acquire knowledge including sentiment information visually.

III. Methodology

In this paper, the analysis is performed according to the process of Figure 1. First, topics are extracted by learning a topic model. Next, we find the sentiment and topic distribution for that topic. Finally, a model is constructed based on these data and this is then analyzed by SEM so that can gain knowledge.

A. Topic Model

The topic model is a method that tries to clarify the structure of a document group by inferring words contained in the topic based on the premise that the document group has a specific topic. In a topic model, a document is a collection of words probabilistically generated by the topic to which it belongs.

Topic models include different methods, such as latent semantic analysis (LSA) [18], LDA [19] and hLDA [20]. The LDA assumes a multi-topic model in which the document is based on mixed topics. LDA has a 1:*n* relationship between documents and topics, not 1:1 like LSA. LDA is considered to be a more natural model in documents, such as review texts that are written in one document about various aspects [19].

HLDA is an extended method of LDA and is a hierarchal model as shown in Figure 2. It has the property of automatically constructing relationships among hierarchical topics. As a learning result, a hierarchical model constructed hierarchically and a keyword group constituting each topic are generated together with their generation probabilities. The specific content of the topic can be inferred from the keyword groups of a topic. In this study, hLDA is used because it is a natural document model and the relationships between topics are defined automatically.

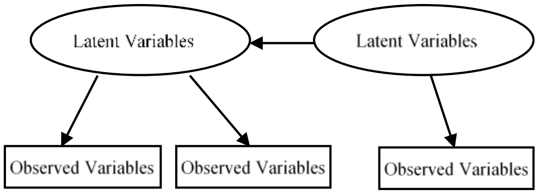


Figure 3. Path model of SEM

B. Sentiment Analysis

Sentiment analysis literally refers to the analysis of sentiments. By using sentiment analysis, such as posted comments, one can determine whether consumers have negative or positive sentiments and the strength of such sentiments. Sentiment analysis can be performed on a per-document or per-sentence basis.

To embed sentiment to SEM explained later, we have to recognize sentiments on each topic for each review. In this study, we regard the average of sentiment values ranging between -1(negative) and 1(positive) as document sentiments by calculating Equation (1) as

$$E_{im} = \frac{1}{|T_i(S_m)|} \sum_{s \in T_i(S_m)} E(s), \quad (1)$$

where E_{im} is the sentiment about the topic T_i of the review R_m; S_m is a set of sentences in R_m and | | is the element number of a set; $T_i(S_m)$ represents the sentence set of S_m, including the topic I; and the function E recognizes the sentiment of a sentence. If there is no sentence related to a topic, the result of Equation (1) is 0 (neural) and regards this sentiment about the topic as neutral. The longer the review, the more likely it is to include other topics. Therefore, it is possible to extract sentiments related to topics more accurately by focusing only on sentences containing topics in reviews.

Here, valence aware dictionary for sentiment reasoning (VADER) [21] is used as function E in the equation. This method is particularly accurate for sentiment analysis in social media. There are several studies that used VADER. One study analyzed the correlation of positive and negative user reviews of mobile apps before and after app update, respectively, by using VADER because VADER has the high precision in the social media field [22]. In VADER, the value of sentiment is represented by -1 to 1 (the closer to -1 the more negative and the closer to 1 the more positive the sentiment). Therefore, the E_{im} outputs the value between -1 and 1.

C. Strucuture Equation Modeling (SEM)

SEM [23] is a method characterized by the use of factor analysis and regression analysis. Factor analysis is the idea that observed variables are based on some hidden factor, and the influence of the factor is to be determined by “correlation” (variance / covariance). Regression analysis is a technique for finding the relationship between a variable to be predicted (target variable) and a variable (explanatory variable, independent variable) that describes the target variable. In other words, SEM can be considered as a factor regression analysis.

The SEM can express causal relationships between variables visually and quantitatively by using a path model, as shown in Figure 3. A path model consists of three elements: latent variables, observed variables, and paths. Latent variables are factors that cannot be observed in actual. Observation variables can actually be observed and are essential for estimating a latent variable. In the path model, latent variables are represented by ellipses and observation variables are represented by rectangles. The causal relationship between such items is represented by the path of the arrow, and the degree of influence is represented by the path coefficient.

D. Construct Path Model and Find Knowledge

Topics that cannot be observed directly are considered as latent variables serving as correspondence between SEM and topic model. The keywords that make up the topic, the sentiment for the topics, and the rating values of each review are the observation variables. From the idea of the topic model that words are generated by topics, each topic is regarded as a factor and the paths from the topics are drawn to the keywords to which the topics are related. Moreover, the paths between topics are drawn from the upper topics to the lower ones based on the idea of the hierarchical structure of the hLDA topics.

Next, we explain the process of incorporating sentiment information into the path model. Sentiment information influences the intention of a model. Thus, we have to carefully determine how to incorporate sentiment information. Generally, sentiments for service are generated as perceived experience (after the service) or the expectation (before using the service). Therefore, the model is expressed by drawing a path to sentiment information from each topic. When we draw a path from the topic to sentiment information, the causal relationship between the sentiment and the topic becomes clear. Moreover, rating evaluation is considered to be generated from the top-level topic that includes all elements. Therefore, by drawing the path from the top-level topic to the rating evaluation, the model can represent the causal relationship with the rating.

Furthermore, by comparing the values of path coefficients from the higher topics to the lower topics, it is possible to find an important factor for the rating. By paying attention to the path coefficient from the lowest topic to the keyword, we can find the degree of influence of more detailed factors. The path coefficients from each topic to sentiment are large and the causal relationship with sentiment could be expressed. By comparing the path coefficient from each topic to sentiment, topics with a larger causal relationship with sentiment can be found.

However, the path model of SEM is usually prone to model identification failure, especially if there are too many latent variables. Conversely, if the number of latent variables is less, the amount of information in the model may be too small for interpretation. As the topic is a latent variable in the path model, the number of topics must also be adjusted. We also need to remove unreliable paths and observation variables with relatively small influence.

IV. Experiments

The purpose of this experiment is to confirm the feasibility of proposed approaches described in Section III. Furthermore, we consider the experimental results.

A. Dataset, Parameters, and Processing

In this analysis, the data must have text data and numerical evaluation data, and it is ideal to have as many review data as possible in order to apply the topic model. In addition, in order to characterize statistical data based on the concept of Bag of Words, the text of one review data must include many words. In this experiment, we employ user-reviews of the datasets published online by Kaggle and Github: the hotel[1], airport[2], app[3] for shops and electronic services[4] for purchasing clothes. Airport, app and electronic services reviews are collected by web scraping. Hotel reviews are provided by Datafiniti's Business Database. Each review has review text with a rating between 1 and 5 or 1 and 10. We also regard a review text as a document. In this method, we have to ensure that the topics and the appearance frequency of the feature words described are included in each document. In addition, we examined reviews of each dataset and understood that a review that passes for a document have about 30 words. Therefore, only documents stated with more than 30 words are used. The app analyzes information from randomly extracted data. The number of reviews after these pre-processing is shown in Table I. In this experiment, sentiments on topics in the lowest level are determined for the construction of a path model. Moreover, T_i in (1) indicates a topic of the lowest level (i.e., topic in third level). Whether a sentence includes or does not include a topic is determined based on whether or not a keyword constituting the topic is included.

As criteria to evaluate the result, we use goodness of fit index (GFI), adjusted GFI (AGFI), root means square error of approximation (RMSEA), and Bayes information criterion (BIC) were used [24][25]. As equations for GFI, AGFI, RMSEA, BIC,

$$GFI = \frac{tr((\sum(\hat{\theta})^{-1}(S - \sum(\hat{\theta})))^2)}{tr((\sum(\hat{\theta})^{-1} - S)^2)}, \tag{2}$$

where $\Sigma(\hat{\theta})$ is the estimated value of covariance matrix and S is value of the actual sample covariance matrix. $tr((A)^2)$ expresses $tr(AA')$,

$$AGFI = 1 - \frac{n(n+1)}{2df}(1 - GFI), \tag{3}$$

where n is the number of observed variables and DF is degrees of freedom,

$$RMSEA = \sqrt{\frac{\max\left[\frac{\chi^2 - DF}{N-1}, 0\right]}{DF}}, \tag{4}$$

where N is the number of samples,

$$BIC = \chi^2 - DF\log(N). \tag{5}$$

And as an equation to calculate degrees of freedom,

$$DF = \frac{1}{2}n(n+1) - p, \tag{6}$$

where p is the number of variables in equation. Equation (2) expresses how well the total variance in the saturation model that includes paths between all possible variables can be explained by the estimation model that is the analysis result of this experiment. A value between 0 and 1 is taken and the closer a value is to 1, the better the model becomes. A value of 0.9 or higher is desirable. GFI is unconditionally improved in fitness as model degrees of freedom decreases. Equation (3) corrects the shortcomings of GFI and penalizes models with many parameters and high complexity. The same value as GFI is taken, and the closer it is to 1 the better the resulting model. If the model is not complex, GFI and AGFI will be close values. Equation (4) is an index that expresses the difference between the model distribution and the true distribution. The fit is good with a value of 0.05 or less, and the fit is bad with 0.1 or more. Equation (5) estimates the posterior probability based on chi-square value when the model is selected. This is used to evaluate the balance between model suitability and the amount of information and is used in carrying out relative evaluation. It is better for the value to be smaller.

In this experiment, we used several packages and libraries: Mallet package [26] for hLDA, Python's nltk package with VADER method [27] for sentiment analysis, and SEM package of R [28] for SEM analysis.

[1] https://www.kaggle.com/datafiniti/hotel-reviews#Datafiniti_Hotel_Reviews.csv
[2] https://github.com/quankiquanki/skytrax-reviews-dataset/tree/master/data
[3] https://www.kaggle.com/usernam3/shopify-app-store#reviews.csv
[4] https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews#

TABLE I. Data and result

Dataset Name	# of Reviews	GFI	AGFI	RMSEA	BIC
Hotel[1]	8104	0.9025	0.8881	0.05525	9188
Airport[2]	13444	0.9152	0.9005	0.05266	12950
App[3]	5442	0.8979	0.8835	0.05960	6848
e-Commerce[4]	19354	0.9213	0.9060	0.05446	19272

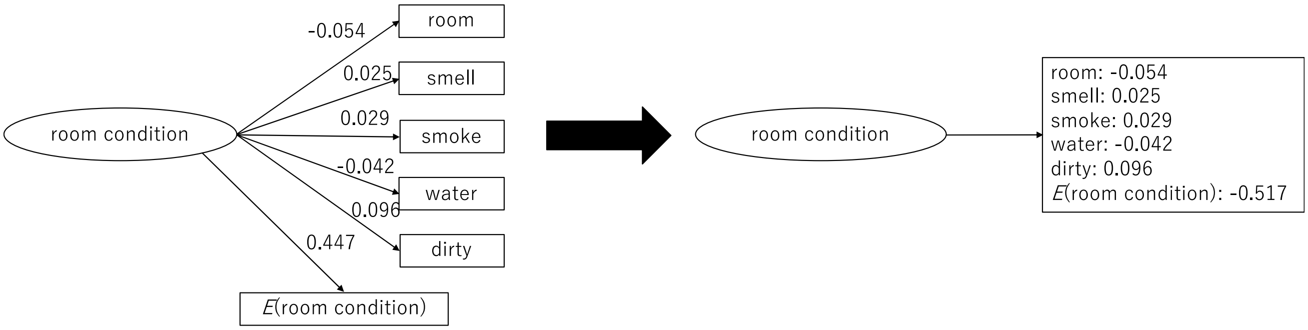


Figure 4. Expression of a path from the latent to the observed variable

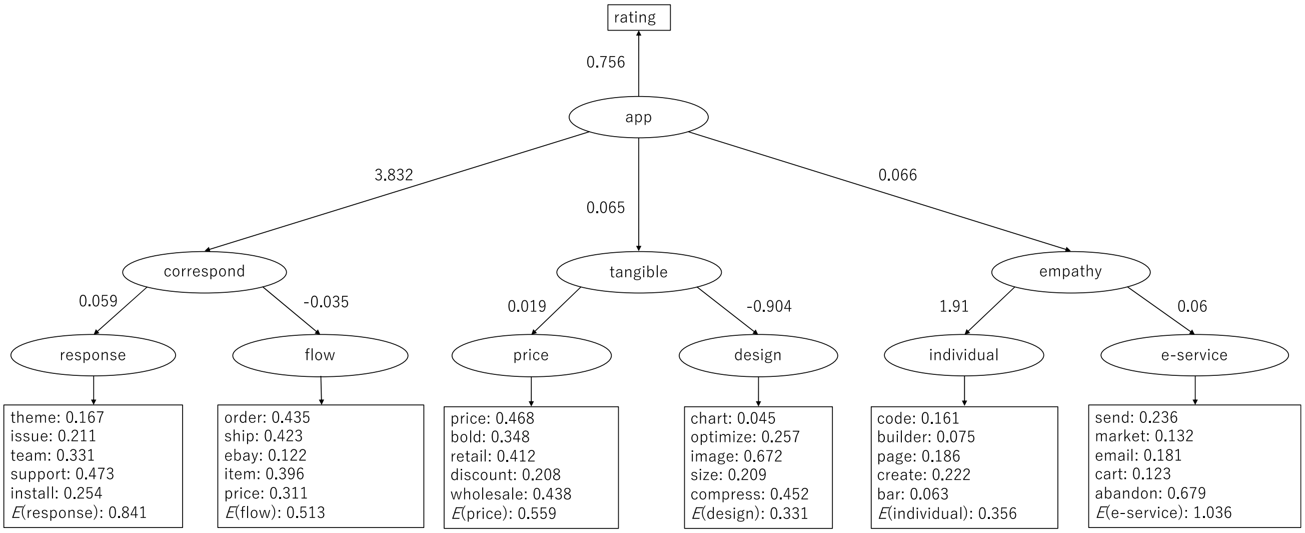


Figure 5. Analysis result of the app dataset

B. Result

Table I shows the calculation results of the evaluation indexes for each data and analyzed models. From Table I, we could find that hotel, airport, and e-commerce models have a GFI of over 0.9 and AGFI maintains high levels. Moreover, none of the models have an of less than 0.05, but it is much closer to 0.5, compared to the model whose fit is bad with 0.1 or more.

It can be said that all of models fit well to the dataset and the constructed models are reliable from the viewpoint of these indices.

As an example, let us show the result of the app dataset in Figure 5. The causal relationship among the keywords that comprise a topic is similar to the depiction in Figure 4. The words at the bottom of the model are those that make up the identified topics from the text data of the review using the topic extraction with hLDA. Here, the topics (latent variables) are estimated by authors from the words that make up each topic. For example, "response" is estimated because it has a large causal relationship with "support" and is considered to be a topic related to responses to actions, such as "install," "team," and "issue." We were able to create a path model based on the hierarchical structure of a text data document group revealed by hLDA. Further, causal relationships can be analyzed by paying attention to arrow and values calculated by SEM between topics or between topics and words or sentiment information at the bottom of the model.

We focus on the "correspond" area with a large path coefficient from the top topic. The "response" is also considered to be an important factor for evaluation because when comparing the two topics under "correspond," the path from "correspond" to "response" has a larger path coefficient. Here, the path between the latent variable "response" and the value of the sentiment "E(response)" has a large coefficient, implying that "response" has a strong relationship with the sentiment strongly. Therefore, it can be considered that the sentiment of "response" also leads to evaluation.

In the same way, when we check the other paths to sentiments, we could find the relationships with and influences to evaluation. From the figure, "response," "flow," "price," and "e-service" have an effect of sentiments (the paths over 0.5) and the "design" and "individual" did not. We are not certain whether the results agree or not, but this specific one indicates which topics lead to emotional satisfaction. In this way, it is possible to improve the service

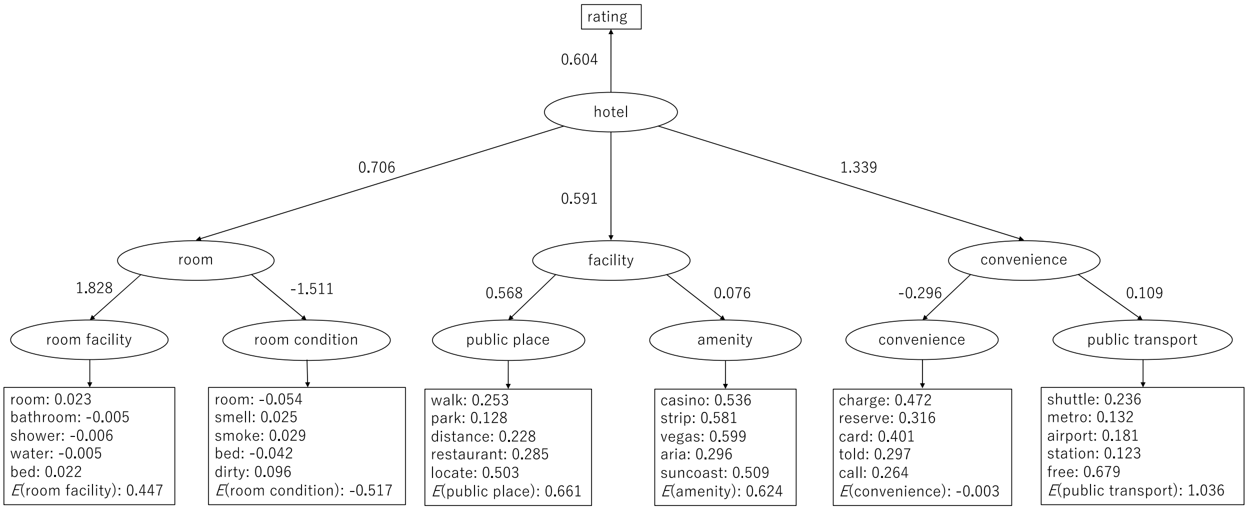


Figure 6. Analysis result of the hotel dataset

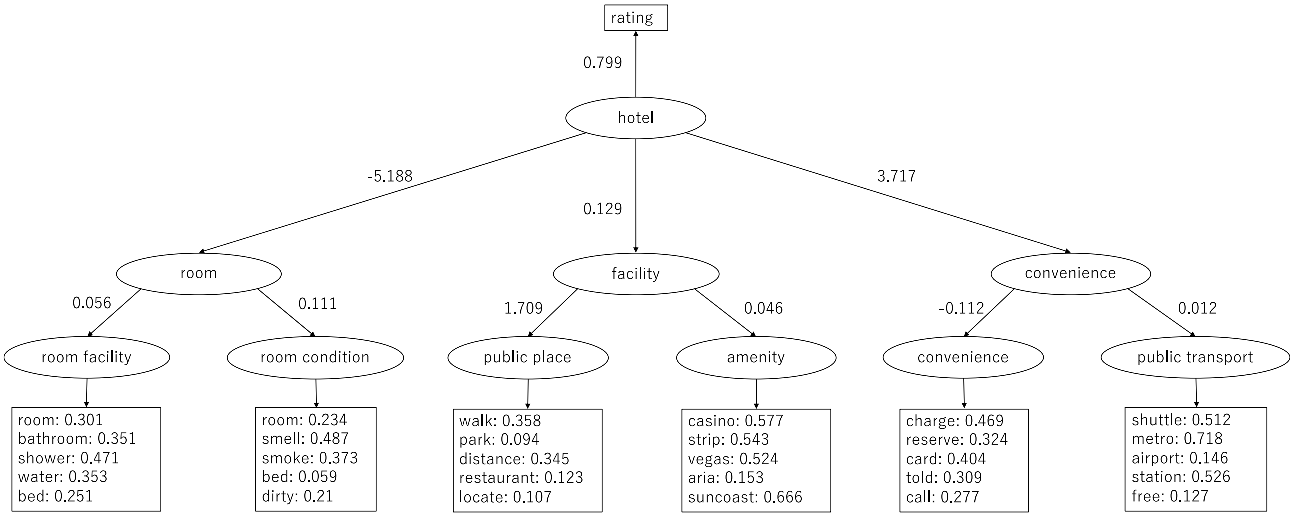


Figure 7. Analysis result of the hotel dataset that deletes sentiment information from Figure 6

by quantitatively understanding the specific service factors that influence to the sentiments and evaluation.

Figure 6 presents another example of analysis. For instance, "room facility" is estimated by different room features namely, "bathroom," "shower," and "bed" and "room condition" is evaluated using "smell," "smoke," and "dirty." The hotel structure can be reviewed by examining the results of the analysis of hotel data. Hotels are evaluated using "room," "facility," and "convenience." By focusing on low hierarchy, the details of the evaluation factors, such as "room condition" and "public transport," can be analyzed. Moreover, the factor that influences sentiment can be comprehensively understood. We focus on the "convenience" area with a large path coefficient because this topic exerts large influence on the evaluation (rating). "Convenience" that has "charge" and "card" has a small effect on sentiment, whereas "public transport" that has "shuttle" and "metro" exerts a large effect. Therefore, "public transport" leads to emotional satisfaction, whereas "convenience" does not.

We then compare the analysis results with and without sentiment information. Figure 7 illustrates the analysis result of the hotel dataset that does not consider sentiment information. We compare Figures 6 and 7. The topic "convenience" composed of "charge" and "reserve" shares a weak relationship with sentiment information. Therefore, even if the sentiment information is deleted, no large difference is observed in the path coefficient between the topic and the words that constitute the topic.

Subsequently, we analyze the topic "public transport," which is strongly related to sentiment information. If sentiment information is not considered, the largest path coefficient is observed in the path to "metro;" otherwise, the path to "free" possesses the largest coefficient This phenomenon occurs because the path coefficient from the path to the words that have strong relationships with sentiment information increases, that is, if "free" has a strong

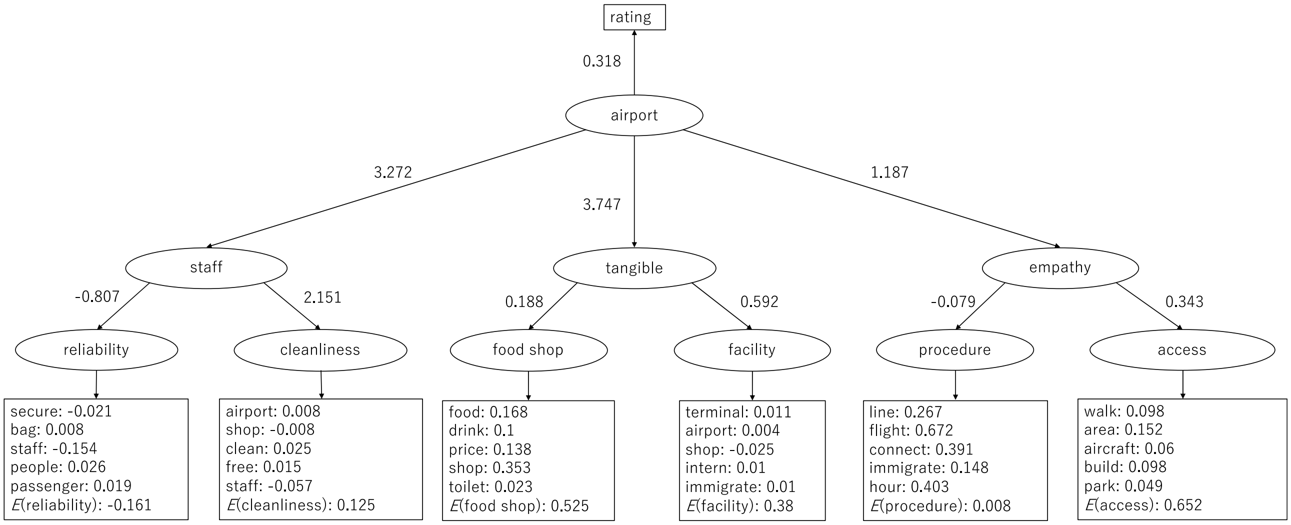


Figure 8. Analysis result of airport dataset

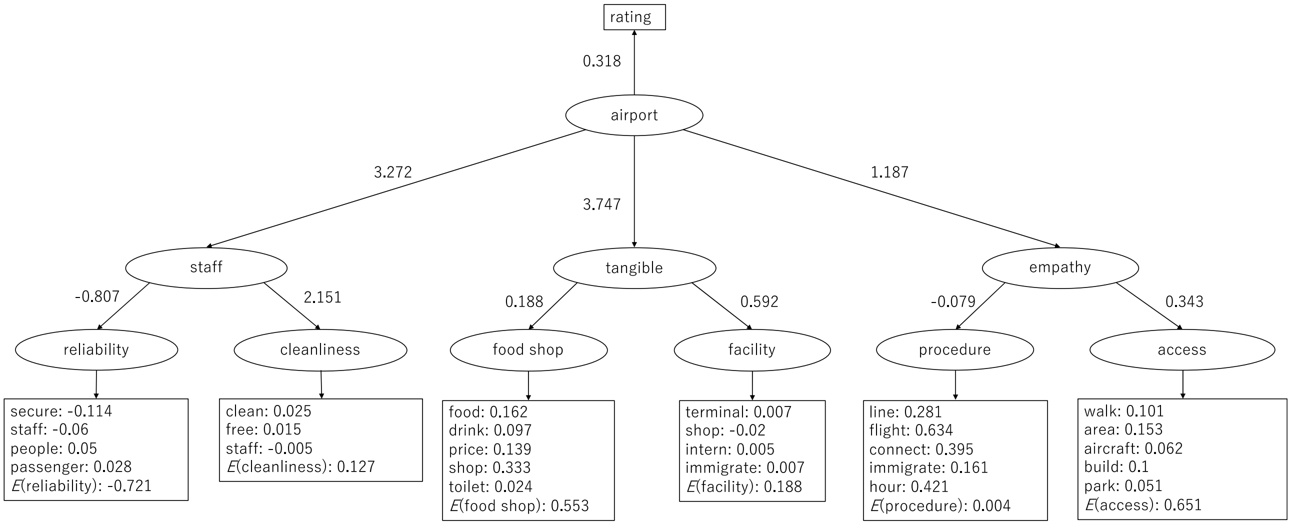


Figure 9. Analysis result of airport dataset that deleted several paths from Figure 8

relationship with sentiment information, then this word is the important factor in the causal model of hotels that considers sentiment information.

In the topic "room facility" in Figure 7, all path coefficients have positive values. However, in the same topic in Figure 6 , which considers sentiment information, paths with negative values, such as those to "smell" and "smoke," appear. This phenomenon can be ascribed to the negative relationship of these words with the sentiment information of the topic "room condition." For instance, the sentiment values in documents that contain these words tend to be negative, whereas those in documents that do not have these words are positive. Therefore, adding sentiment information to the topic leads to the clarification of the negative factor.

TABLE II. EVALUATION INDICES OF AIRPORT DATASET

Figure Name	GFI	AGFI	RMSEA	BIC
Figure 7	0.9152	0.9005	0.05266	12950
Figure 8	0.9210	0.9007	0.05275	11358
Figure 9	0.9275	0.9134	0.05287	9885

In summary, a causal model that considers sentiment information can be constructed.

This study aims to improve the interpretability of the causality model. Figure 8 shows the result of the airport dataset. The figure displays several paths that have small path coefficients. A causality model with enhanced interpretability can be constructed by deleting these paths because the two variables connected by a small path coefficient have almost no causal relationship. Figure 9 displays the result after deleting the paths in Figure 8 that have small path coefficients (path coefficient < 0.01), except those with sentiment information. Figure 10 shows the result after deleting the paths in Figure 9 that have small path

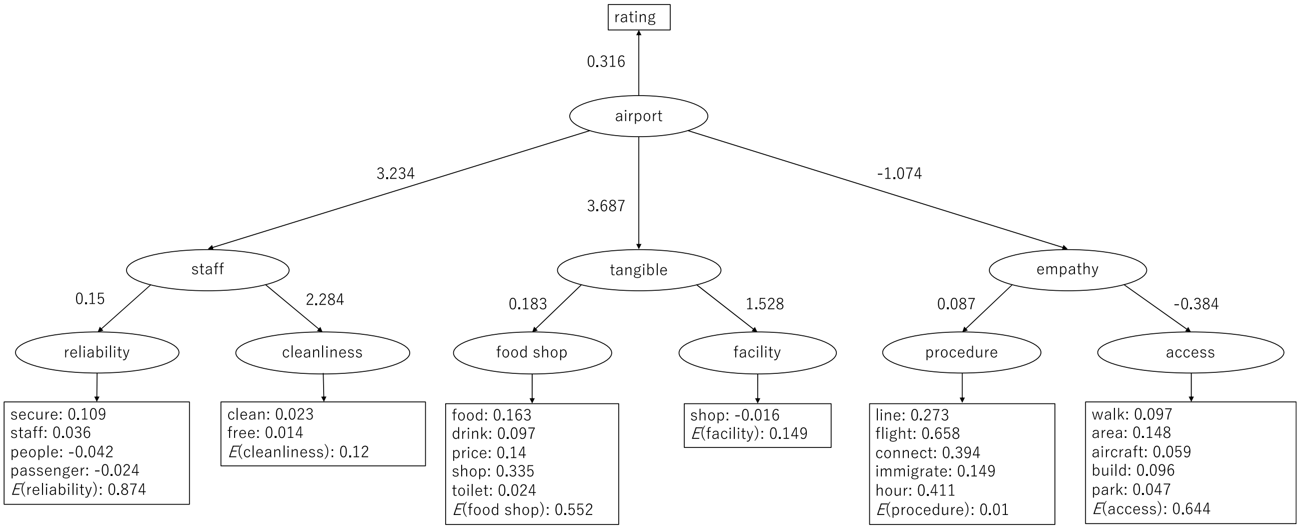


Figure 10. Analysis result of airport dataset that deleted several paths from Figure 9

coefficients (path coefficient < 0.01), except those with sentiment information.

Table II summarizes the calculation results of the evaluation indices for Figures 7, 8, and 9. GFI and AGFI increase as the paths that have small path coefficients are deleted, whereas RMSEA decreases because of the change in the number of observed variables. The results suggest that all figures fit well to the dataset, and the constructed models are reliable from the viewpoint of these indices.

An easy-to-interpret causality model of the airport dataset can be constructed from Figure 10 because the paths that have small causal relationships are deleted. In other words, the amount information decreases, but we can construct a simple model and focus only on the important elements.

V. Discussion and Future work

From the experiment, we found that sentiment information is useful for analyzing services, but we have to consider improving sentiment expression. For example, we extracted sentiment information of topics based on (1), but this equation does not consider the length of the sentence. Nevertheless, it enables us to accurately determine the sentiment on the topic by considering the weight based on the sentence length. For example, longer sentences are more likely to include other topics. Therefore, it may be possible to extract sentiments related topics more accurately by reducing the impact of such sentences on sentiments of specific topics. Secondly, when two or more topics are included in one sentence, even if it is used in a contrasting sentence, such as "(Text about TOPIC A) but (Text about TOPIC B)," the same sentiment value is calculated for the topic. If there is a conjunction (e.g., "but"), a more accurate sentiment analysis can be performed by further processing, such as dividing. Thirdly, several factors such as "smells" in Figure 4 are considered negative but it would be positive for several people. Therefore, it is possible to express this situation by dividing reviewer into a group that thinks the factor is negative and a group that thinks factor is positive and expressing it to path model. Finally, in this paper, the accuracy improvement and knowledge are obtained by constructing path models under different assumptions during the construction of the path model.

Furthermore, we consider the hierarchical topic structure to construct the path model. In this study, we use hLDA to extract such structure. Several methods can be used to extract the hierarchical topic structure. Zhu et al. proposed an extraction method [29] that combines a biterm topic model (BTM) [30] and Bayesian rose trees (BRTs) [31]. The present study extracts the topics by using BTM and constructs a hierarchical structure by utilizing BRTs. Moreover, this study adopts simBRT to account topic similarity. Viegas et al. proposed CluHTM [32], which is a novel non-probabilistic hierarchical topic modeling strategy based on non-negative matrix factorization and CluWords [33]. This method ensures topic coherence and reasonable topic hierarchies and uses the utilization as an original cross-level stability analysis metric to define the number of topics and the shape of the hierarchical structure. The abovementioned methods can be used to accurately estimate the document structure.

A topic is defined as a bag of words without explicit semantics. In this study, the contents of the topics are estimated using the words that compose them. However, the topic model loses objectivity. To address this issue, we can use topic labeling. Several methods can be used to add semantic labels to the topic model. Nalasco et al. proposed an automatic labeling technique by using a new candidate selection algorithm and three scoring methods [34]. Bhatia et al. proposed a neural embedding approach that involves automatic topic labeling by using Wikipedia article titles [35]. Mao et al. proposed an automatic labeling technique for hierarchical topic structures [36].

VI. Conclusion

In this paper, we analyzed the causal relationships in service by using SEM and sentiment information. We constructed the path model by using hLDA and sentiment analysis between topics and sentiments. The findings of the experiment using the user reviews of airports, hotels, shopping apps, and electronic services show the feasibility of our proposed model. We summarize the following findings from the experiments:

- We obtained knowledge by analyzing service while considering sentiments.
- We determined the impact on the rating of each topic.
- We obtained the causal relationship between each topic and sentiment quantitatively and provided clues for further analyses.

Service analysis that considers sentiment information is conducted by this study. We found that sentiment information has the relationship with service evaluation.

We also performed service analysis considering sentiment and obtained knowledge reflecting sentiment information from the user reviews. The consideration of sentiment information is essential for service analysis, and the creation of path models with sentiment information is considered effective in extracting information that helps increase service satisfaction. It is suggested that the analysis process in this paper may provide useful knowledge for service analysis and service improvement. On the one hand, this can be used by service providers in improving services and creating new services. Service providers can quantitatively find factors that have major impacts on the evaluation of services and customer sentiments. On the other hand, it can be used by service users to efficiently grasp the outline of services that are not formed. Although we analyzed the indefinite service in the experiments, it can be applied to other things like tangible products. The potential applicability is high because analysis is performed from the text.

References

[1] T. Ogawa and R. Saga. “Text-based Causality Modeling with Emotional Information Embedded in Hierarchic Topic Structure”, Proceedings of the Ninth International Conference on Social Media Technologies, Communication, and Informatics, pp. 15-20, November 2019.

[2] Central Intelligence Agency: *The World Factbook: GDP - Composition, by sector of origin*. [Online]. Available from: https://www.cia.gov/library/publications/the-world-factbook/fields/214.html, [retrieved: October, 2019].

[3] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, “SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality”, Journal of Retailing, vol. 64, No. 1, pp. 12-40, 1988.

[4] Airports Council International. ACI: *Airport Service Quality, ASQ. : The ASQ Barometer*. [Online]. Available from: https://aci.aero/customer-experience-asq/services/asq-barometer/, [retrieved: October, 2019].

[5] V. Liljander and T. Strandvik, “Emotions in service satisfaction”, International Journal of Service Industry Management, vol. 8, Issue 2, pp. 148-169, 1997.

[6] J. J. Cronin and S. A. Taylor, “Measuring Service Quality: A Reexamination and Extension”, Journal of Marketing, vol. 56, No. 3, pp. 55-68, July 1992.

[7] M. Ali and S. A. Raza, "Service quality perception and customer satisfaction in Islamic banks of Pakistan: the modified SERVQUAL model", Total Quality Management & Business Excellence, vol. 28, Issue 5-6, pp. 559-577, November 2015.

[8] P. K. Sari, A. Alamsyah, and S. Wibowo, "Measuring e-commerce service quality from online customer review using sentiment analysis", Journal of Physics: Conference Series, vol. 971, Issue 1, 2018.

[9] F. D. Orel and A. Kara, “Supermarket self-checkout service quality, customer satisfaction, and loyalty: Empirical evidence from an emerging market”, Journal of Retailing and Consumer Services, vol. 21, Issue 2, pp, 118-129, March 2014.

[10] R. Kunimoto and R. Saga, “Causal Analysis of User’s Game Software Evalution Using hLDA and SEM”, IEEJ, vol. 135, Issue 6, pp. 602-610, 2015.

[11] A. M. Al-Mhasnah, F. Salleh, A. Afthanorhan, and P. L. Ghazali, “The relationship between services quality and customer satisfaction among Jordanian healthcare sector”, Management Science Letters, vol. 8, Issue 12, pp. 1413-1420, 2018.

[12] N. Aidin, “Revisiting customers’ perception of service quality in fast food restaurants”, Journal of retailing and consumer services, vol. 34, pp. 70-81, Janualy 2017.

[13] K. E. Zannat, D. R. Raja, and M. S. G. Adnan, “Pedestrian Facilities and Perceived Pedestrian Level of Service (PLOS): A Case Study of Chittagong Metropolitan Area, Bangladesh”, Transportation in Developing Economies, vol. 5, Issue 2, pp. 1-16, April 2019.[D]

[14] G. R. Bivina and M. Parida, “Modelling perceived pedestrian level of service of sidewalks: a structure equation approach”, Transport, vol. 34, No. 3, pp. 339-350, May 2019.

[15] S. Rahi, M. A. Ghani, and F. M. Alnaser, “The Influence of E-Commerce Services and Perceived Value on Brand Loyalty of Banks and Internet Banking Adoption: A Structural Equation Model (SEM)”, The Journal of Internet Banking and Commerce, vol. 22, No. 1, pp. 1-18, April 2017.

[16] K. Lee and C. You, “Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews”, Journal of Air Transport Management, vol. 71, pp. 28-44, August 2018.

[17] L. Martin-Domingo, J. C. Martín, and G. Mandsberg, “Social media as a resource for sentiment analysis of Airport Service Quality(ASQ)”, Journal of Air Transport Management, vol. 78, pp. 106-115, July 2019.

[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis”, Journal of The American Society for Information Science, vol. 41, Issue 6, pp. 391-407, 1990.

[19] D. M. Blei, A. Y. Ng, J. B. Edu, and M. I. Jordan, “Latent dirichlet allocation”, The Journal of Machine Learning Research, No. 3, pp. 993-1022, 2003.

[20] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process”, Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 17-24, 2003.

[21] C. J. Hutto and E. Gilbert, “VADER: a parsimonious rule-based model for sentiment analysis of social media text”, International AAAI Conference on Web and Social Media, 2014.

[22] L. Xiaozhou, Z. Zheying, and S. Kostas, “Sentiment-aware Analysis of Mobile Apps User Reviews Regarding Particular Updates”, Proceedings of The Thirteenth Intetnational Conference on Software Engineering Advances, pp. 99-107, 2018.

[23] C. J. Anderson and W. D. Gerbing, “Strucutural equation modeling in practice: A review and recommended two-step approach”, Psychologcal Bulletin, vol. 103, No. 3, pp. 411-423, May 1988.

[24] K. Schermelleh-Engel, H. Moosbrugger and H. Müller, “Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures”, Methods of Psychological Research Online, vol. 8, No. 2, pp. 23-74, 2003.

[25] G. Schwarz, “Estimating the Dimension of a Model”, Annanls of Statistics, vol. 6, No. 2, pp. 461-464, 1978.

[26] A Kachites, “Mallet: A Machine Learning for Language Toolkit”, http://mallet.cs.umass.edu, [retrieved: October, 2019].

[27] “Natural Language Toolkit – NLTK 3.4.4 document”, https://www.nltk.org, [retrieved: October, 2019].

[28] R. Ihaka, R. C. Gentleman, “The R Project for Statistical Computing”, https://www.r-project.org, [retrieved: October, 2019].

[29] J. Zhu, X. Li, M. Peng, J. Huang, T. Qian, et al., “Coherent Topic Hierarchy: A Strategy for Topic Evoluationary Analysis on Microblog Feeds”, Proceedings of 16th International Conference on Web-Age Information Management, pp. 70-82, June 2015.

[30] X. Yan, J. Guo, Y. Lan and X. Cheng, “A biterm topic model for short texts”, WWW’13: Proceedings of the 22nd international conference on World Wide Web, pp. 1445-1456, May 2013.

[31] C. Yee, Y. W. The and K. A. Haller, “Bayesian Rose Trees”, UAI’10: Proceedings of the Twenty-Sixth Conference on Uncertainly in Artificial Intelligence, pp. 65-72, July 2010.

[32] F. Viegas, W. Cunha, C. Gomes, A. Pereira, L. Rocha, et al., “CluHTM-Semantic Hierarchical Topic Modeling based on CluWords”, Proceedings of the 58th Annual Meeting of the Association for Computation Linguistics, pp. 8138-8150, July 2020.

[33] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, et al., “CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling”, WSDM’19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Minig, pp. 753-761, January 2019.

[34] D. Nolasco and J. Oliveira, “Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data”, Proceeding of the 49th Hawaii International Conference on System Science (HICSS), pp. 358-367, 2016.

[35] S. Bhatia, J. H. Lau and T. Baldwin, “Automatic labeling of topics with neural embeddings”, Proceedings of the 26th International Conference on Computing Linguistics: Human Language Technologies, pp. 1536-1545, 2011.

[36] X. L. Mao, Z. Y. Ming, Z. J. Zha, T. S. Chua, H. Yan and X. Li, “Automatic labeling hierarchical topics”, ACM International Conference Proceeding Series, pp. 2383-2386, 2012.

Agent-Based Simulation of Strain and Motivation in Human Work Performance

Stephanie C. Rodermund
Business Informatics I
Trier University
Behringstraße 21, 54296 Trier, Germany
rodermund@uni-trier.de

Bernhard Neuerburg
German Aerospace Center (DLR)
Linder Höhe, 51147 Köln, Germany
bernhard.neuerburg@dlr.de

Fabian Lorig
Internet of Things and People Research Center (IoTaP)
Malmö University
Nordenskiöldsgatan 1, 211 19 Malmö, Sweden
fabian.lorig@mau.se

Ingo J. Timm
Business Informatics I and
German Research Center for Artificial Intelligence
SDS Branch Trier (Cognitive Social Simulation)
Trier University
Behringstraße 21, 54296 Trier, Germany
itimm@uni-trier.de

***Abstract*—Even though the relevance of the "human factor" on the performance of work processes is well known, the design and optimization of such processes, e.g., in factories, often strongly focuses on machines. Especially intrinsic mental states such as *strain* and *motivation* can influence the human workers' performance and thus the organizational outcome. This paper is based on a previous agent-based model of human work processes and extends this model using Atkinson's theory of achievement motivation. The combination of the job demands-resources model with a more advanced motivation theory allows for a more sophisticated and realistic modeling of task selection based on its difficulty, individual competencies, and perceived attractiveness. Experiments are presented, to demonstrate the model's capability to simulate human work performance and the mutual influences between job demands, resources, personal resources, as well as the intrinsic mental states of strain and motivation.**

***Keywords*–Human Work Performance; Agent-based Modeling; Job Demands-Resources Model; Strain; Achievement Motivation.**

I. Introduction

In previous work, the relevance and impact of the "human factor" on the performance of work processes has been outlined and a model for the simulation of human work performance based on strain and motivation has been proposed [1]. This is relevant, as peoples' workplaces are constantly changing, especially as digitalization progresses, and as we believe that this digital revolution should be oriented towards employees' needs. Yet, people often subordinate to IT systems and thus disempower themselves [2]. For example, a scheduling system in a call center distributes incoming calls without considering individual needs of the call center agents. Consequences are not only physical but also psychological strains like burn-out.

Digital transformation should not be rejected in general as it has the potential to make work processes more efficient. Current approaches for designing and optimizing work processes, e.g., the production of goods in a factory, often make use of simulation and focus on machine processes. Examples are predictive maintenance or throughput time optimization. Here, downtimes of machines or queuing strategies are analyzed to identify optimal process configurations. In reality, however, human workers can also influence the performance of such production processes, e.g., due to unavailability, distraction, or overload. Existing frameworks for the analysis of industrial service provision processes often neglect the human factor and only allow for the modeling and simulation of machines in production lines.

In a production plant, human workers may be assigned a series of orders with different difficulties to be processed during the working day. The workers' performance can be measured by the ratio of completed orders in relation to the total number of orders. While machines do not show performance fluctuations when being confronted with an immense workload or time pressure, human workers tend to be susceptible to such influences. Intrinsic processes of motivation and strain are driving factors influencing their performance [3]. Still, during the planning and implementation of work processes, human beings are often only considered as workforces without individual intrinsic needs, even though their significance and importance are well known, e.g., modeling of humans in Business Process Model and Notation (BPMN). To achieve a more adequate integration of humans into these processes as well as to increase performance and organizational outcome, individuals and their intrinsic needs must be represented individually and realistically.

Based on these considerations, the authors of this article have developed an agent-based model of human work performance by utilizing the Job Demands-Resources model (JDR model), which includes motivation and strain as intrinsic mental states [1]. They investigated the agents' performance in a simple work context in which orders of various difficulties need to be completed in a limited time. In different simulation experiments, plausible results were generated that confirm the mutual influence of motivation and strain.

This paper adapts the previous model and presents two main extensions that focus on the definition of motivation by using Atkinson's motivation theory [4] as well as on the impact of strain on the duration of order processing using a Performance Moderator Function (PMF) [5]. This allows for a more sophisticated and realistic modeling of individual task selection based on the tasks' difficulty, individual competencies, and the subjectively perceived attractiveness of tasks. To model workers and their behavior, Agent-Based Modeling (ABM) and especially the Belief-Desire-Intention (BDI) architecture of practical reasoning [6] are used, which are established in modeling of human cognitive decision-making [7]–[10].

The article is structured as follows. In Section II, related work on the field of modeling strain and motivation in ABM is presented and discussed. In this regard, the concept of performance motivation is introduced, which serves as a theoretical basis for extending motivation in the proposed model. Furthermore, the flexible Job Demands-Resources model is introduced, which is well-established in psychology and investigates factors in the working environment that may lead to burn-out, especially focusing on those factors causing a stressful situation and mental effort for the worker [11]. Subsequently, an extended agent-based model of work performance is introduced in Section III. In Section IV, the results of simulation experiments are discussed to analyze the model's adequacy to represent human work performance. Finally, Section V provides a summary as well as an outlook on future work.

II. Background

There are several frameworks for modeling and optimizing industrial processes, e.g., Enterprise Dynamics or Anylogic [12], which strongly focus on functionalities of machines in manufacturing. These frameworks lack in the representation of human resources such that the "human factor" cannot be considered properly when measuring the overall performance. However, other areas, e.g., the representation of social networks, lay emphasis on an adequate representation of human beings. Here, agent-based models that utilize sociological and psychological behavioral theories are well-established [13]–[15]. This article introduces an extended agent-based model of human work performance including the intrinsic processes of strain and motivation, which in future work could be used to represent workers in existing frameworks. In the following, we discuss existing work on agent-based models including stress and motivation formation and present the psychological JDR model, that serves as the basis for our implementation.

A. Modeling and Simulation of Strain

In ABM, various approaches exist that include psychological strain in behavioral development. Silverman's generic agent architecture contains a working memory (BDI decision logic) and four subsystems: Physiological System, Emotive System, Cognitive System and Motor/Expressive System [16]. In the strictly modularized approach, the calculation of an integrated stress value is part of the Physiological System, which is defined as a function of exhaustion, time pressure, and event strain. Fatigue is represented via available physiological resources and time pressure results from perceived stimuli. Event strain is the result of negative emotions of the Emotive System [17]. Based on these variables, different coping strategies are initiated using a PMF. Silverman proposes an inverted-U shaped PMF, which was first introduced by Janis & Mann and has since been replicated and validated several times. Depending on the integrated stress value, different coping strategies are chosen: Unconflicted Adherence and Change, Vigilance, Defensive Avoidance, and Panic. This PMF is characterized by an activating effect of stress on performance in addition to the limiting effects [5]. Duggirala et al. apply this conceptual model in an agent-based simulation of strain at work [18]. They selected the variables *task arrival volume*, *pending tasks*, and *work hours* to calculate the integrated strain value and to determine the coping strategies. However, by choosing work hours for determining exhaustion, they have missed Silverman's consideration of individual resources.

Ashlock and Cage also simulate strain at work using an agent-based model and a strain factor consisting of individual strain tolerance and number of stressors [19]. Still, strain is difficult to quantify and validate, especially using static mathematical formulas that are limited to a number of variables. For this reason, Morris et al. investigated system dynamics of strain to model agents by representing strain as causal loop diagram and stock-flow diagram [20]. In the BDI extension BRIDGE, strain is, similar to Silverman's approach, part of the implicit behavior and only influences the deficiency needs and overrules selected intentions [21]. Another broad research field, in whose models strain is also considered, (e.g., [22]), is crowd simulation. Strain influences behavior generation mainly reactively, but this is due to the frequent application context of emergency evacuations, where deliberative behavior is less important.

Most models include two aspects: Firstly, the models focus on stimuli during the genesis of strain and secondly in doing so, they neglect the consideration of resources that can significantly reduce the amount of strain generated. Such models do not recognize strain as the result of intrinsic processes although psychology has already sufficiently shown the degree to which cognitive processes occur regarding strain for a long time (e.g., [23]).

B. Modeling and Simulation of Motivation

In ABM, when considering motivation as part of the decision-making process, models can be distinguished by the motivations' directionality, i.e., whether motivation is caused by external factors or if it is merely generated intrinsically by the individual. Maslow's hierarchy of needs as an intrinsically oriented motivation theory, e.g., is implemented by Spaiser and Sumpter [24] as well as Silverman [16]. In these models, the agent's actions focus primarily on covering deficiency and growth needs, and mostly neglect environmental influences on motivation development. As mentioned above, the BRIDGE architecture also uses this theory to define an agent's goals and desires [21]. Using Vroom's extrinsically oriented expectation theory, the agent's decision making is modeled on the basis of its expected subjective value of a future event in his environment [25] [26].

Following Atkinson's concept of achievement motivation [4], behavior is aimed at the self-assessment of a competency, in confrontation with a standard of quality that one wishes to achieve or exceed [27, p.59]. Achievement motivation is affected both by external tendencies T_{ex} (e.g., striving for reward or avoiding punishment) and internal tendencies T_i, which result from the conflict of hope for success $T_s = M_s \cdot W_s \cdot A_s$ and fear of failure $T_f = M_f \cdot W_f \cdot A_f$, where

- $M_s(M_f)$ represents the success (failure) motive (stable disposition of a person, describing the capability to experience pride when having success (M_s) and shame when being unsuccessful (M_f)),
- $W_s(W_f)$ is the subjective expectancy of success (failure) (a person's expectancy that an action leads to an anticipated goal (or not); this variable changes due to experience), with $W_s + W_f = 1$, and
- $A_s(A_f)$ is the incentive of a success (failure) (a person's pride exceeds with difficulty of a given task) [27, pp. 59].

Individuals with a motive profile of $M_s > M_f$ are *success-oriented*, which means that they tend to look for goals that they want to achieve. These are achieved by minimizing the difference between the current status and the goal status. In contrast to this, a motive profile of $M_s < M_f$ means that these individuals are *failure-oriented*. They tend to avoid failure by maximizing the distance between the current status and the goal status [28]. Atkinson also states that the incentive for success can be described as $A_s = 1 - W_s$ (cf. [4, p. 94]) and, thus, solely depends on the subjective expectancy of success. This is based on the assumption that accomplishing a task that appears to be very difficult and, therefore, probably not achievable is perceived more attractive than an easily accomplishable task [4, p. 94]. A similar thought applies to the incentive for failure A_f. If an individual defines a task as easy to accomplish with a high value of W_s, the shame and embarrassment felt by the individual is also high in case the accomplishment of this task fails. Therefore, the incentive of failure can be described as $A_f = -W_s$. This leads to an adaption of the resulting tendency to $T_r = (M_s - M_f) \cdot (P_s - P_s^2)$. Among other things, e.g., the persistence in completing a task [29] [4, pp. 110], achievement motivation can be used to explain the selection of tasks of various degrees of difficulty [4, p. 99].

Achievement motivation has, so far, only been used in a few agent-based models. For instance, Merrick and Shafi (2013) investigated the effect of the three motive profiles of achievement, power, and affiliation motivation in situations of several mixed motive games. The authors demonstrate that the perception of the agents differs from each other according to their current motive profile composition [30]. Di Pietrantonio et al. developed an agent-based model of organizational work performance based on both the workers' abilities as well as their motivational needs [32]. Therefore, they also make use of the *Three Needs Theory* [31], which includes the motive profiles of achievement, affiliation, and power motivation. The authors investigate the effect of different motive profile distributions and the workers' own abilities while working in teams on the overall performance, which is defined as the number of completed tasks after a specific number of time steps [32]. To the authors' knowledge, Atkinson's achievement motivation model is only sparsely used in ABM. Among just a few others, Merrick [33] uses this motivation theory. She utilizes an experiment from human psychology and simulates it with agents to prove the suitability of the concept for use in an agent-based model.

The introduced approaches for ABM of motivation mainly rely on subjectively perceived environmental factors and largely neglect the mutual influence of intrinsic factors, e.g., between perceived strain and motivation, although the relation between these factors has already been described, e.g., by Dignum et al. [21].

A well-known model that both considers stressors (stimuli), resources, and the influence of motivation, is the JDR model by Demerouti et al. [11]. The JDR model is an empirically evaluated model that has been flexibly used in a variety of scenarios such as to predict job burn-out [34], organizational commitment [35], connectedness [36], and work engagement [37]. The model consists of two processes: a health impairment process and a motivational process (see Figure 1). The health impairment process is concerned with how job demands affect individual strain. Job demands can be stressors like workload, emotional demands, or organizational changes [38].

As part of the motivational process, job resources are main predictors for motivation and engagement. While job demands consume energetic resources and cause strain, job resources fulfil basic psychological needs and generate motivation. Thus, job demands and resources initiate two different processes but these processes are not independent because job resources can buffer the impact of job demands on strain and job demands can reduce the generation of motivation through job resources (see Figure 1). Due to these moderation effects, there is also a direct relationship between strain and motivation. By using the model, predictions can be made about employee well-being, job-performance, and respectively the aggregated performance of a company.

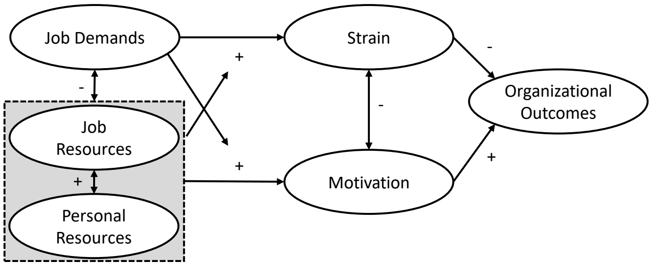


Figure 1. Job Demands-Resources Model [39].

The model was extended several times by the authors, in particular to include job crafting and self-undermining, and was transferred into a theory based on several meta-analyses [3], [40]. In this work, one of the first extensions of the model is used to significantly reduce the complexity of the simulation and to focus on the prediction of job performance [39].

III. An Agent-Based Model of Work Performance

In this section, an extended agent-based model of human work performance is introduced that combines the BDI architecture and the JDR model presented in Section II. The workers are modeled based on the BDI architecture of practical reasoning [6], which organizes goals (desires), information about the environment and the own conditions (beliefs), and action-oriented measures (intentions) into mental states. To this end, we also make use of the JDR model presented in Section II. By utilizing both models, a strict modularization is achieved, which can be easily extended and exchanged by other theories and models.

Figure 2 shows the basic concept of the agent-based model of human work performance. Following the JDR model, the agent's environment consists of sets of $JobDemands$, $JobResources$, and $PersonalResources$ that impact internal processes forming $strain$ (α) and $motivation$ (ζ). These, in turn, determine the agent's action as well as the corresponding duration of the action and, thus, the organizational outcomes. Here, this is equal to the individual performance.

Referring to the factory example introduced in Section I, the agent is confronted with a set of $Orders$ that is composed of the two sets $UnfinishedOrders$ and $FinishedOrders$ (Equation (1)). Initially, $|Orders|$ is equal to $|UnfinishedOrders|$. If an order $i \in UnfinishedOrders$ is completed, it is deleted from this set and added to $FinishedOrders$. Each of the orders has a certain difficulty $diff_i \in \mathbb{N}$, which is defined within a range of set difficulties. The difficulty of an order expresses how much time is required to execute it. As job demands represent stressors like workload (see Section II), $difficulties$ is introduced, which represents the agent's workload on one working day. It is composed of the sum of difficulties $diff_i$ for each $i \in UnfinishedOrders$ (Equation (2)).

$$Orders = FinishedOrders \bigcup UnfinishedOrders \qquad (1)$$

$$difficulties = \sum_{i=1}^{|UnfinishedOrders|} diff_i \qquad (2)$$

A working day is defined by a number of time steps $totalTime \in \mathbb{N}$, where $t \in \mathbb{N}$ represents the current time that has already elapsed. At each time step, the $remainingTime$ to complete all $UnfinishedOrders$ is computed (Equation (3)). The difficulty level corresponds to the minimum number of time units required to process an order and depends on the agent's $skillRank \in \mathbb{N}$, i.e., its work-related know-how. A lower value of $skillRank$ means that less time units are needed to complete one difficulty level. The $skillRank$ together with the overall $remainingTime$ to complete all orders form the agent's set of $JobResources$.

The agent's set of $PersonalResources$ is comprised of its general motives $motiveSuccess \in \mathbb{N}$ and $motiveFailure \in \mathbb{N}$ as well as its own $selfEfficacy \in \mathbb{R}$. The motives are based on Atkinson's achievement motivation model introduced in Section II, that is used as the underlying motivation theory. $selfEfficacy$ represents the subjectively perceived competence to perform actions effectively [41] [42]. The agent's $PersonalResources$ can be gathered from an input of empirical data (see Section V).

$$remainingTime = totalTime - t \qquad (3)$$

Job demands initiate a health impairment process that affects the agent's individual strain. Job resources, on the other hand, have a moderating effect on strain and buffer the impact of the job demands. $strain$ (Figure 2, Function α) represents the experienced pressure as the ratio between the unfinished orders $difficulties$ and the $remainingTime$ to complete them (Equation (4)).

$$\alpha\colon strain = \frac{difficulties}{remainingTime} \qquad (4)$$

$Motivation$ is formed in a process that is influenced by job resources, job demands, and personal resources. Based on the achievement motivation introduced in Section II, we require the two motives $motiveSuccess$ and $motiveFailure$ as well as the subjective probability of success to define motivation for this model. In [1], motivation is defined as the general and objective probability that "represents whether the agent is able to perform the open orders in the given time based on its own $skillRank$ at time t". As this definition does not yet take into account individual motives and subjective probabilities, it is used to represent the objective probability of success at time t $objProb_t$ (Figure 2, Function β) (see Equation (5)). As $objProb_t$ represents a probability, its value is normalized to the interval [0,1]. A higher value of this variable implies that the agent is objectively capable of completing the whole set of unfinished orders in the remaining time.

$$\beta\colon objProb_t = \frac{remainingTime}{skillRank_t \cdot difficulties} \qquad (5)$$

The subjective probability of success for a specific order difficulty $subjProbS_{diff} \in$ [0,1], on the one hand, is composed of a general and objective probability $objProb_t$. The subjective component of $subjProbS$ is introduced by the agent's $selfEfficacy \in$ [0,1], which defines the agent's own conviction of being able to complete tasks of high complexity [42]. Nicholls [43] states that this reflects Atkinson's assumption that the "degree of difficulty can be inferred from the subjective probability of success" [28, p.362]. Furthermore, the influence of $selfEfficacy$ on an agent's performance reduces with increasing task complexity [41]. Thus, the agent's $subjProbS$ is represented by the decay of $selfEfficacy$ based on the objective probability to complete all remaining orders and referring to the level of difficulty of the respective order (see Equation (6)).

$$subjProbS_{diff} = selfEfficacy^{(1-objProb_t)\cdot diff} \qquad (6)$$

Consequently, the subjective probability of success is used to define $motivation$ (Figure 2, Function ζ) for each remaining order difficulty $diff$ as follows:

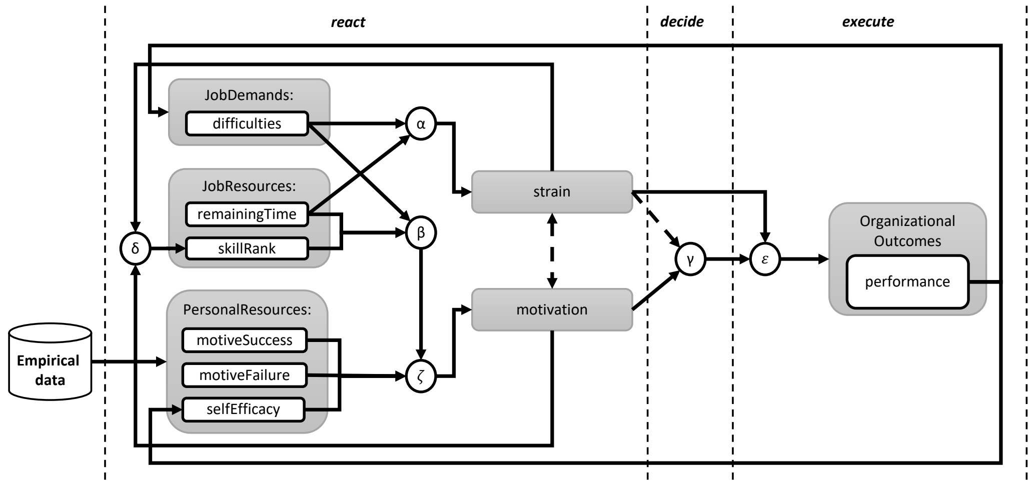
$$\zeta\colon motivation_{diff} = (motiveSuccess - motiveFailure)\cdot$$
$$(subjProbS_{diff} - subjProbS^2_{diff}) \qquad (7)$$

For the purpose of simplicity, we neglect the external tendency T_{ex} for now, since we assume a controlled environment without an external reaction as reward or punishment to the work done. As the next task to accomplish (Figure 2, Function γ), the agent always selects the difficulty of available orders for which the highest motivation value $motivation_{diff}$ exists (see Equation (8)).

$$\gamma\colon \arg\max_{diff} motivation_{diff} \qquad (8)$$

$Strain$ and $motivation$ represent an agent's set of $IntrinsicStates$. Both values are normalized to $[0, 1]$, relative to the minimal and maximal possible values of the variables.

To calculate the agent's productivity, and the time the agent needs to complete a task, an inverted-U shaped PMF



```
Algorithm: BDI ControlCycle

1   B := initializeBeliefs (input)
2   D := initializeDesires (input)
3   I := {}
4   action := none
5   repeat
6       B := brf (B, D, action)
7       I := decide (B, D, action)
8       action := selectIntention (I)
9       execute (action, B)
10  end repeat
```

Figure 2. Job demands-Resources Model as Agent-Based Model (left) and Algorithm (right).

(Figure 2, Function ε) is introduced following Silverman's approach described in Section II. It considers both the limiting effect and the activating effect of stress on performance. Depending on the current strain value, the agent can behave according to five different coping strategies (see Figure 3), which determine the number of ticks required to complete an order.

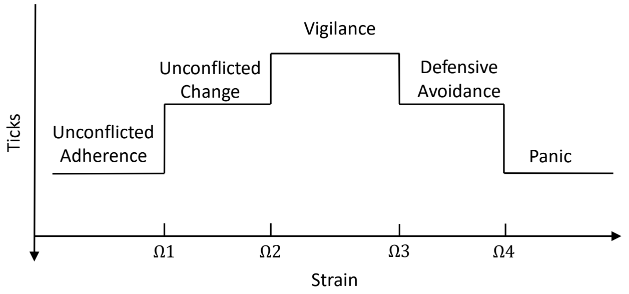


Figure 3. Performance Moderator Function: Inverted-U shaped [17]

The strain thresholds Ω1 to Ω4 are derived from Silverman's work [17]. The required number of ticks ($ticks$) is calculated using following Function ε depending on the default number of ticks ($ticks_{def}$), which an agent at least needs to fulfil a given order:

$$\varepsilon\colon ticks = \begin{cases} ticks_{def} \cdot 1.3, & strain \in [0.0, 0.1] \\ ticks_{def} \cdot 1.15, & strain \in]0.1, 0.25] \\ ticks_{def}, & strain \in]0.25, 0.75] \\ ticks_{def} \cdot 1.15, & strain \in]0.75, 0.9] \\ ticks_{def} \cdot 1.3, & strain \in]0.9, 1.0] \end{cases} . \quad (9)$$

Following the example introduced in Section I, $performance$ is measured using the ratio of $FinishedOrders$ to the overall number of $Orders$ (Equation (10)).

$$performance = \frac{|FinishedOrders|}{|Orders|} \quad (10)$$

The algorithm in Figure 2 shows the BDI control cycle that determines the agent's behavior formation process. First, the internal states as well as a variable determining the next action to perform are initially set (lines 1-4). Based on the general BDI architecture, the agent's behavior in our model is formed by passing various phases that consider and construct the mental states. These can be divided into *react*, *decide*, and *execute* (see [44]). In *react* (*belief-revision-function (brf)*), the agent processes perceived information and updates its beliefs (B) about the current situation and intrinsic states. In *decide*, based on the updated beliefs and the agent's desires (D), the agent updates its intentions (I). Considering these, an action to perform next is chosen, before it is carried out in *execute*.

The agent's beliefs B are composed of the four sets $JobDemands$, $JobResources$, $PersonalResources$, and $IntrinsicStates$ (see Equation (11)). Based on the beliefs B that are generated and updated in *react*, the agent decides for an unfinished order to proceed with next, to reach its sole desire, i.e., completing all orders.

$$\begin{aligned} B = {} & JobDemands \bigcup JobResources \bigcup \\ & PersonalResources \bigcup IntrinsicStates \\ \Rightarrow B = {} & \{difficulties, remainingTime, skillRank, \\ & motiveSuccess, motiveFailure, \\ & selfEfficacy, strain, motivation\} \end{aligned} \quad (11)$$

In the *decide* phase, the agent decides for a difficulty $diff$ of orders it wants to process next. For this, the agent computes motivation values for each remaining difficulty and decides for a difficulty with the highest motivation and, thus, for the intention I to commit to (see Figure 2, Function γ). Consequently, *decide* is only processed if the current order has been completed in the preceding time step. The chosen difficulty (I) is used to pick the next order ($action$) to complete, which is then performed in *execute*. Starting from the initial value, the $skillRank$ adapts in dependence to the values of $motivation$ and $strain$ (decrease or increase of value) and to the current order's difficulty (strength of decrease or increase of value) (Figure 2, Function δ). Furthermore, based on the expected time needed to complete an order in comparison to the actual time that it takes for the agent, the value of *selfEfficacy* is modified. If the agent is performing as expected (defined by thresholds $\Omega 2$ and $\Omega 3$) the value is slightly increased and if the productivity strongly deviates it is slightly decreased, so if strain is transcending the thresholds $\Omega 1$ and $\Omega 4$ (see, e.g., [42]). After each time step t, the *performance* is used to update the orders' difficulties.

IV. Simulating Work Performance: Experiments and Results

In this section, the agent-based model of work performance is evaluated based on a case study and compared to previous simulation results from [1]. First, the main findings from the initial paper are presented. Then, the simulation setup for this article's model is defined and the additional model input variables are specified. Finally, the findings are presented and the assumptions derived from these are discussed.

This article's model presents an extension of the agent-based model of work performance defined in [1]. The authors specified the agent's initial *skillRank*, the *difficultyRange*, which represents the range of difficulties, orders in the experiment can have and the available *timeCapacity* (see Table I). Furthermore, the number of *Orders* is set to 20 and the maximum value of the agent's *skillRank* is fixed at 10. After 30 replications of each defined experiment, Figure 4 shows the results separated by the variation of *timeCapacity*, whereas the x-axis depicts the initial input value of the variable *skillRank*. The y-axis shows the performance of the agent. The boxplots' colors represent the orders' *difficultyRange*, darkgrey represents a range of 1-3, lightgrey for 1-5, and white for a range of 3-5.

TABLE I.
SCENARIO SPECIFICATION ORIGINAL EXPERIMENT [1].

timeCapacity		*difficultyRange*		*skillRank*
smallTimeCapacity		1-3		1
suitableTimeCapacity	×	1-5	×	5
highTimeCapacity		3-5		10

The authors discuss three main findings as well as several observations from the experiment results:

1) An increasing *timeCapacity* leads to increased performance: In a scenario with a *high timeCapacity*, the agent is capable to complete all or a majority of orders in the given time, without considering the respective *skillRank*.
2) A low *skillRank* does not equal a high performance: The performance is represented via the ratio of finished orders. Agents with a *skillRank* of 1 tend to choose orders of a high difficulty and, thus, finish less orders in summary because of the adaption in *skillRank* after a bad performance and the respective *strain* and *motivation*.
3) A *difficultyRange* of 3-5 leads to the worst performance: Thus, the mean performance throughout the simulation runs is 0.52, whereas ranges 1-3 and 1-5 lead to mean values of 0.69 and 0.63. This leads to the conclusion that a balanced order compilation is more purposeful as it, on the one hand, demands the worker enough to keep his interest and, on the other hand, allows for phases of lower concentration while completing orders of a low difficulty level [45].

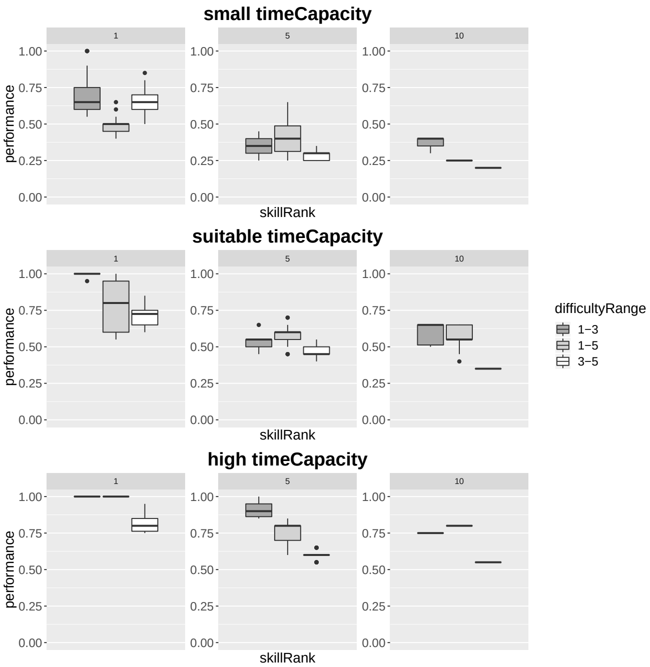


Figure 4. Experimental results in the original experiment [1]. Performance depending on *timeCapacity*, *skillRank*, *difficultyRange*.

Furthermore, the authors address some exceptions to their main findings, namely:

- In *small timeCapacity* and *skillRank* $= 1$ the performance is worse for the order difficulties in a range of 1-5 as for 3-5,
- in *small timeCapacity* and *skillRank* $= 5$ as well as in *suitable timeCapacity* and *skillRank* $= 5$ or 10 the performance is worse for the order difficulties in a range of 1-3 as for 1-5 and
- a *skillRank* of 10 leads to extreme performance measures without outliers.

The first two exceptions are explained by the way *strain* and *motivation* as well as choosing a next order difficulty are defined in the model. In both exceptions, the agent chooses high difficulties first which, caused by the progressing time, leads to increasing *strain* and decreasing *motivation* and ultimately to less finished orders. The third exception is due to a low *motivation* value resulting from the high *skillRank* as well as the restriction of the model to generate a higher *skillRank* than 10. With decreasing *remainingTime*, the *strain* value increases and the *skillRank* is not allowed to improve.

A. Simulation Setup

In this article, the agent-based model of work performance from [1] is extended by making use of a specific motivation theory, the achievement motivation defined by Atkinson as well as the effect of *strain* on the ticks needed to complete an order (see Section III). To be able to compare the simulation outcomes of the extended model and the basic model, further variable specifications need to be mentioned.

The adapted model in Section III introduces the set of $PersonalResources$ as an additional input of the JDR model. The set composes of the variables $motiveSuccess$, $motiveFailure$ and $selfEfficacy$. To include these parameters, the scenario specification in Table I needs to be adapted. Following [46], a possible way to obtain these person-specific motives is a questionnaire containing of 10 items, which can take values of 1 to 5 each. Here, five items refer to the motive for success and five belong to the motive for failure. Thus, $motiveSuccess$ and $motiveFailure$ each can take values in the interval [0,20]. In the scenarios defined in this experiment, these variables vary in steps of five, leading to a set of [5, 10, 15, 20]. Equally, the value for $selfEfficacy$ can be derived with a questionnaire (see e.g., [47]), and in this model can take values between 0.25 to 1.0 in steps of 0.25. Accordingly, 1728 experiments are defined ($timeCapacity$ (3) × $difficultyRange$ (3) × $skillRank$ (3) × $motiveSuccess$ (4) × $motiveFailure$ (4) × $selfEfficacy$ (4) = 1728). Additionally, the value of $ticks_{def}$, that is needed in PMF (see Equation (9)) to determine the productivity, is set to the agent's current $skillRank$, as this variable was defined as the number of ticks needed to complete one difficulty of an order. Because the model includes stochastic processes each experiment is repeated 30 times.

B. Simulation Results and Discussion

The simulation results in Figure 5 show the experimental results separated by $timeCapacity$. As in Figure 4, the x-axis shows the initial input of $skillRank$ and the y-axis shows the output of the agent's performance. The boxplots separate by color in the three available difficulty ranges 1-3 (darkgrey), 1-5 (lightgrey) and 3-5 (white). The overall tendencies described earlier in this section remain for the adapted model presented here, too: With an increasing $timeCapacity$ the agent's performance increases. Hence, the mean performance value increases for $skillRank$ of 10 and $difficultyRange$ of 3-5 from 0.16 in $small\ timeCapacity$ to 0.47 in a scenario with a $high\ timeCapacity$. Second to that, the $difficultyRange$ of 3-5 leads to the worst performances of a mean value of 0.42, whereas ranges of 1-3 and 1-5 lead to performance means of 0.73 and 0.64.

Besides these general tendencies, the experiment output shows some deviations from the initial paper. As stated above, the $difficultyRange$ affects the performance in a way that high difficulties (3-5) lead to the worst performances. In contrast to the findings in [1], this effect is present in each scenario separated by $timeCapacity$ and $skillRank$. A reason can be found in the definition of $motivation$ of the original model that is ultimately dependent on the input parameters (e.g., $remainingTime$). Based on that calculation of $motivation$, in some scenarios the agent always chooses the highest difficulty available. In comparison, $motivation$ here is extended by the personal motive profile of the agent. Agents with a higher $motiveSuccess$ as $motiveFailure$ tend to decide for orders with a medium probability of success (e.g., in a range of 1-5 the difficulty 3 is predominantly chosen), whereas an agent with a higher value of $motiveFailure$ than $motiveSuccess$ results in choosing border options (difficulties that are very likely or very unlikely to complete successfully). This leads to a higher distribution in the choice for options for each of the defined $difficultyRanges$ and thus to a decrease in performance from the ranges 1-3 over 1-5 to 3-5.

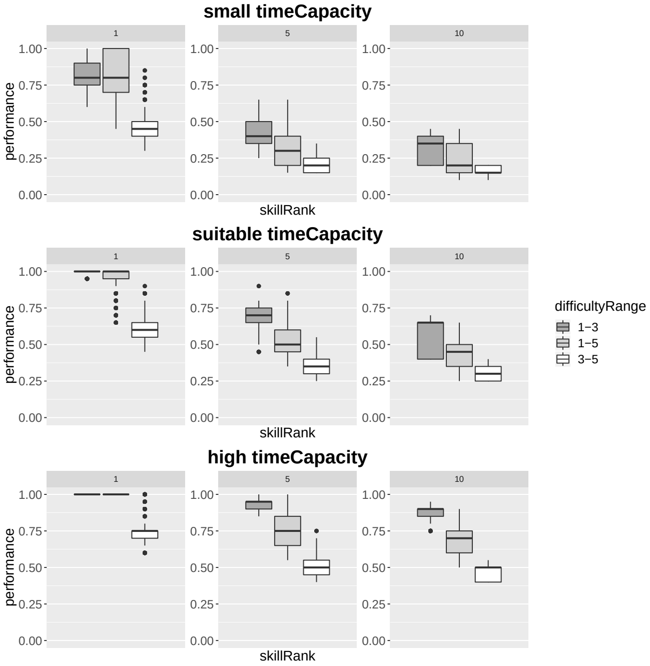


Figure 5. Performance depending on $timeCapacity$, $skillRank$, $difficultyRange$.

Finally, a $skillRank$ of 10 produces a less uniform picture than in the original experiments. The values of the minimal and maximal performances of these agents span a wider value range of at a maximum 0.3 in a $suitable\ timeCapacity$ and $difficultyRange$ of 1-3. Overall, a $skillRank$ of 10 still produces the worst performances in each scenario, but especially in $high\ timeCapacity$, the comparison of the resulting performance of the current model and the one in [1] shows an increase of performance of 0.2 at a maximum. As is defined in the scenario specification at the beginning of this section, the maximum $skillRank$ is set to 10. Therefore, this value can not deteriorate due to the agent's poor performance. On the contrary, the skill of an agent can be improved based on a decreased $strain$ and increased $motivation$ value. Furthermore, the presence of different motive profile distributions leads to a higher spread in a choice for difficulties. Additionally, agents with an equally distributed motive profile randomly choose one of the orders, regardless of the respective difficulties [4, p.99]. As each of these decisions influences the agent's overall performance, due to the adapting variables $strain$ and $motivation$, the observed behavior can be explained.

To investigate the effect of different motivation profile distributions, Figure 6 shows the agent's performance (on the y-axis) separated by the available $timeCapacity$ (x-axis). The expressions *HighLow*, *HighHigh*, *MediumMedium*, *LowLow* and *LowHigh* refer to the composition of the agent's motive profile in the sequential order $motiveSuccess$ followed by $motiveFailure$. *HighLow* means that the agent under investigation has a high value of $motiveSuccess$ (here: 20), while $motiveFailure$ has a low value, e.g., of 5 (cf. Table II). The two motive profiles *HighHigh* as well as *LowLow* are not completely equally distributed. This is based on the fact that equal values completely negate the effect of the respective other, which leads to a complete random selection of

TABLE II.
SPECIFICATION OF MOTIVE PROFILES.

Motive profile	$motiveSuccess$	$motiveFailure$
HighLow	20	5
HighHigh	20	15
MediumMedium	10	10
LowLow	5	10
LowHigh	5	20

difficulties. Therefore, in deviation from the experiments in, e.g., [33] or [48] a fifth motive profile *MediumMedium* was added, that represents this equally distributed motive profile. A profile defined as *HighHigh* is thus characterized as having a maximum value of $motiveSuccess$ and the second highest value of $motiveFailure$. For the corresponding profile *LowLow* the value of $motiveFailure$ is set to the higher value in comparison to $motiveSuccess$.

Throughout the simulation runs, a motive profile of *HighLow* shows the best performance results with an overall mean of 0.63. An agent's best performance can be found at $high\ timeCapacity$ with a mean value of 0.79 and a maximum of 1. In all scenarios, this motive profile is capable of reaching a maximum performance by completing all available orders. An agent with a high $motiveSuccess$ and a low $motiveFailure$ tends to choose a medium order difficulty (with a medium probability of success), which could lead to a relatively constant value of $strain$, since the progressing time is neither very large nor very small. This, in turn, influences the time needed for the next order defined by the PMF.

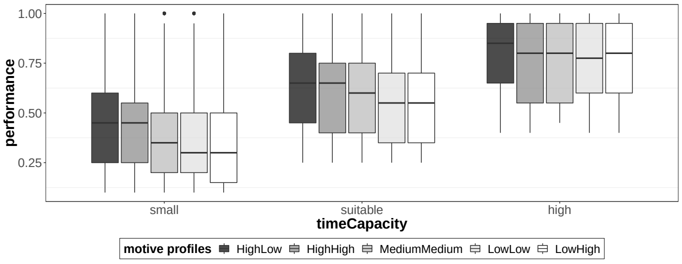


Figure 6. Performance depending on motivation profiles.

In contrast to this, the profile *LowLow* produces the worst performance with a mean of 0.58. Compared to the motive profile *LowHigh*, which is often at a similar level, the mean value only slightly differs from it with a distance of 0.01 at $high\ timeCapacity$ (*LowLow*: 0.76 and *LowHigh*: 0.77). *LowHigh* leads to extreme decisions due to the high proportion of $motiveFailure$. Hence, in situations with time pressure, the $strain$ value either increases because the agent decides for a difficult order that demands a long processing time or slightly decreases because the agent chooses the other extreme with an easy and less time consuming order. This may explain the wider output space for this profile in $small\ timeCapacity$ in contrast to *LowLow*. On the other hand, in a scenario with enough time ($high\ timeCapacity$) an agent with a *LowLow* profile chooses order difficulties more randomly, which can lead to less finished orders. For the profile *LowHigh* the agent more probably relies on an order difficulty of 3, as with decreasing time the subjective probability of success might shift to this difficulty as time progresses (see Equation (6)).

An agent that has a high $motiveSuccess$ as well as $motiveFailure$ possesses the second-highest performances throughout the presented scenarios, whereas the mean performance values duplicate those of *HighLow* in a small as well as suitable $timeCapacity$. With such a profile, the agent chooses more randomly but with a shift towards the kind of decision-making of *HighLow* due to $motiveFailure = 15$ rather than 20 (as it is the case in *MediumMedium*).

The motive profile *MediumMedium* neglects the impact of the two motives as they neutralize the impact of each other (see Equation (7)). This agent chooses a difficulty based on a random manner. This leads to a medium overall performance of 0.60 and a mean of 0.78 in $high\ timeCapacity$. Here, the performance is just as high as with a profile of *HighHigh*.

V. Conclusion and Future Work

In this article, an extended agent-based model of human work performance is presented that makes use of the JDR model and was based upon the model presented in [1]. A decision-behavior based on the general BDI architecture was introduced and adapted to the processes defined in the JDR model including a representation of strain and motivation as well as the mutual influences of job resources, job demands, personal resources and intrinsic mental states. The motivation is based upon a theoretical foundation of Atkinson's achievement motivation and extended the definition of the original model mentioned before. Within several experiments, the impacts of the input variables $timeCapacity$, $skillRank$, and $difficultyRange$ on the overall performance of the agents were analyzed. Furthermore, the impact of different motive profiles was investigated. The experimental results revealed that the model is capable of producing realistic working performance including intrinsic processes of strain and motivation. The extension of the original model by achievement motivation and PMF allows for a more sophisticated and realistic representation of performance. Hence, different motive profile distributions lead to a decision behavior similar to empirical findings in literature [4, p.99].

In future work, we plan on conducting empirical experiments with workers in a controlled working environment (see, e.g., [49]). In these experiments, we aim at identifying stressors and resources and measure individual reactions like strain, especially by biosignals (see, e.g., [49] [50] [51]). Additionally, Atkinson's achievement motivation relies on three general determinants, whereas one of them (incentives A_s and A_f) can be fully represented by the probability of success. The motive of success as well as the motive of failure can be measures by using a *revised Achievement Motive Scale (AMS-R)* [46]. Furthermore, the general self-efficacy of a person can be measured using the *Allgemeine Selbstwirksamkeitskurzskala (ASKU) (General Self-efficacy Short scale)* [47]. To measure the individually perceived workload of human test persons, the *NASA-TLX* test could be used [52]. By using these measurement scales, the subjectively perceived situation of the respective test person can be included in the model.

Furthermore, we need to improve the existing model in several respects. The model shows the best results for orders within difficulty range 1-3. As discussed in Sec. IV-A, a varied order difficulty should lead to best performances, due to a balanced ratio of exertion and relaxation [45]. To receive a more realistic representation, the effects of missing challenges could be included. A difficulty range of 1-3 would thus theoretically lead to a worse performance than a range of 1-5. The agents' performance should be measured by showing how much of the workload has been completed. Thus, not only the proportion of finished orders, but the difficulties of the finished orders should be taken into account, too. Additionally, the effect of $motivation$ as well as the motive profiles on persistence could be investigated [4, pp.110ff] [29]. In this context, the effect of orders that are not fully or incorrectly made could be examined. Furthermore, working in teams should be included in the model. This could result in improved organizational outcomes as poor performances of some members may be offset by good performances of others by the interaction.

VI. Acknowledgements

We would like to thank the German Aerospace Center (DLR) and Internet of Things and People Research Center (IoTaP) at Malmö University for facilitating the writing of this paper. The work has not been done in collaboration with or for these institutions.

References

[1] S. C. Rodermund, B. Neuerburg, F. Lorig, and I. J. Timm, "Simulating Strain and Motivation in Human Work Performance: An Agent-Based Modeling Approach Using the Job Demands-Resources Model." in Proceedings of the Eleventh Conference on Advances in System Simulation (SIMUL 2019), Valencia, Spain, 2019, pp. 8–13.

[2] H. Kagermann, "Chancen von Industrie 4.0 nutzen (Seizing Opportunities of Industry 4.0)," in Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung · Technologien · Migration, T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2014, pp. 603–614.

[3] A. B. Bakker and E. Demerouti, "Job Demands–Resources Theory," Wellbeing: A complete reference guide, 2014, pp. 1–28.

[4] J. W. Atkinson and D. Birch, "An Introduction to Motivation (Rev. Ed.)," New York: Van, 1978.

[5] M. F. Rice, "Reviewed Work: Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment by Irving L. Janis, Leon Mann." The Annals of the American Academy of Political and Social Science, vol. 449, no. 1, 1980, pp. 202–203.

[6] A. S. Rao and M. P. Georgeff, "Bdi Agents: From Theory to Practice." in ICMAS, vol. 95, 1995, pp. 312–319.

[7] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," Proceedings of the National Academy of Sciences, vol. 99, no. Supplement 3, May 2002, pp. 7280–7287.

[8] W. Jager and M. Janssen, "The Need for and Development of Behaviourally Realistic Agents," in Multi-Agent-Based Simulation II, G. Goos, Hartmanis, J., van Leeuwen, J., S. Sichman, J., F. Bousquet, and P. Davidsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2581, pp. 36–49.

[9] J. O. Berndt, S. C. Rodermund, and I. J. Timm, "Social Contagion of Fertility: An Agent-based Simulation Study," in Proceedings of the 2018 Winter Simulation Conference (WSC). Gothenburg, Sweden: IEEE, Dec. 2018, pp. 953–964.

[10] L. Reuter, J. Berndt, and I. Timm, "Simulating Psychological Experiments: An Agent-Based Modeling Approach," in Proceedings of the Fourth International Conference on Human and Social Analytics (HUSO 2018), Wilmington, DE, USA, 2018, pp. 5–10.

[11] E. Demerouti, A. Bakker, F. Nachreiner, and W. Schaufeli, "The job demands-resources model of burnout." Journal of Applied Psychology, vol. 86, no. 3, 2001, pp. 499–512.

[12] E. Serova, "The Role of Agent Based Modelling in the Design of Management Decision Processes," The electronic journal information systems evaluation, vol. 16, no. 1, 2013, pp. 71–80.

[13] M. W. Macy and R. Willer, "From Factors to Actors: Computational Sociology and Agent-Based Modeling," Annual review of sociology, vol. 28, no. 1, 2002, pp. 143–166.

[14] T. Balke and N. Gilbert, "How Do Agents Make Decisions? a Survey," Journal of Artificial Societies and Social Simulation, vol. 17, no. 4, 2014, pp. 13–.

[15] E. R. Smith and F. R. Conrey, "Agent-Based Modeling: A New Approach for Theory Building in Social Psychology," Personality and social psychology review, vol. 11, no. 1, 2007, pp. 87–104.

[16] B. G. Silverman, "More Realistic Human Behavior Models for Agents in Virtual Worlds: Emotion, Stress, and Value Ontologies," University of Pennsylvania/ACASA Technical Report, Tech. Rep., 2001.

[17] B. Silverman, M. Johns, J. Cornwell, and K. O'Brien, "Human Behavior Models for Agents in Simulators and Games: Part I: Enabling Science with PMFserv," Presence, vol. 15, 04 2006, pp. 139–162.

[18] M. Duggirala, M. Singh, H. Hayatnagarkar, S. Patel, and V. Balaraman, "Understanding Impact of Stress on Workplace Outcomes Using an Agent Based Simulation," in Proceedings of the Summer Computer Simulation Conference, 2016.

[19] D. Ashlock and M. Page, "An agent based model of stress in the workplace," in 2013 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). IEEE, 2013, pp. 114–121.

[20] A. Morris, W. Ross, and M. Ulieru, "A system dynamics view of stress: Towards human-factor modeling with computer agents," in 2010 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2010, pp. 4369–4374.

[21] F. Dignum, V. Dignum, and C. M. Jonker, "Towards Agents for Policy Making," in International Workshop on Multi-Agent Systems and Agent-Based Simulation. Springer, 2008, pp. 141–153.

[22] Y. Mao, S. Yang, Z. Li, and Y. Li, "Personality trait and group emotion contagion based crowd simulation for emergency evacuation," Multimedia Tools and Applications, 2018, pp. 1–28.

[23] R. S. Lazarus and S. Folkman, Stress, Appraisal, and Coping. Springer publishing company, 1984.

[24] V. Spaiser and D. J. T. Sumpter, "Revising the Human Development Sequence Theory Using an Agent-Based Approach and Data," Journal of Artificial Societies and Social Simulation, vol. 19, no. 3, 2016.

[25] A. Sharpanskykh, "Modeling of Agents in Organizational Context," in Multi-Agent Systems and Applications V, H.-D. Burkhard, G. Lindemann, R. Verbrugge, and L. Z. Varga, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4696, pp. 193–203.

[26] A. Sharpanskykh and S. H. Stroeve, "An agent-based approach for structured modeling, analysis and improvement of safety culture," Computational and Mathematical Organization Theory, vol. 17, no. 1, Mar. 2011, pp. 77–117.

[27] F. Rheinberg, "Grundriss der Psychologie, Band 6, Motivation (Outline of Psychology, Volume 6, Motivation)," Kohlhammer Urban-Taschenbücher, Stuttgart, vol. 6, 2006.

[28] J. W. Atkinson, "Motivational Determinants of Risk-Taking Behavior." Psychological review, vol. 64, no. 6p1, 1957, p. 359.

[29] N. T. Feather, "The study of persistence." Psychological bulletin, vol. 59, no. 2, 1962, p. 94.

[30] K. E. Merrick and K. Shafi, "A Game Theoretic Framework for Incentive-Based Models of Intrinsic Motivation in Artificial Systems," Frontiers in psychology, vol. 4, 2013, p. 791.

[31] M. Brewster Smith, "The Achieving Society by David C. McClelland," History and Theory, vol. 3, no. 3, 1964, pp. 371–381.

[32] J. Di Pietrantonio, R. M. Neilan, and J. B. Schreiber, "Assessing the impact of motivation and ability on team-based productivity using an agent-based model," Computational and Mathematical Organization Theory, vol. 25, no. 4, 2019, pp. 499–520.

[33] K. E. Merrick, "A Computational Model of Achievement Motivation for Artificial Agents," in AAMAS, 2011, pp. 1067–1068.

[34] A. B. Bakker, E. Demerouti, and M. F. Dollard, "How job demands affect partners' experience of exhaustion: Integrating work-family conflict and crossover theory." Journal of Applied Psychology, vol. 93, no. 4, 2008, pp. 901–911.

[35] A. B. Bakker, M. Van Veldhoven, and D. Xanthopoulou, "Beyond the Demand-Control Model: Thriving on High Job Demands and

Resources," Journal of Personnel Psychology, vol. 9, no. 1, 2010, pp. 3–16.

[36] K. A. Lewig, D. Xanthopoulou, A. B. Bakker, M. F. Dollard, and J. C. Metzer, "Burnout and connectedness among Australian volunteers: A test of the Job Demands–Resources model," Journal of vocational behavior, vol. 71, no. 3, 2007, pp. 429–445.

[37] A. B. Bakker, J. J. Hakanen, E. Demerouti, and D. Xanthopoulou, "Job resources boost work engagement, particularly when job demands are high." Journal of educational psychology, vol. 99, no. 2, 2007, pp. 274–284.

[38] A. B. Bakker and E. Demerouti, "The Job Demands-Resources Model: State of the Art," Journal of managerial psychology, vol. 22, no. 3, 2007, pp. 309–328.

[39] D. Xanthopoulou, A. B. Bakker, E. Demerouti, and W. B. Schaufeli, "The role of personal resources in the job demands-resources model." International journal of stress management, vol. 14, no. 2, 2007, pp. 121–141.

[40] A. B. Bakker and E. Demerouti, "Multiple levels in job demands-resources theory: implications for employee well-being and performance," Handbook of well-being, 2018.

[41] A. D. Stajkovic and F. Luthans, "Self-efficacy and work-related performance: A meta-analysis." Psychological bulletin, vol. 124, no. 2, 1998, p. 240.

[42] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change." Psychological review, vol. 84, no. 2, 1977, p. 191.

[43] J. G. Nicholls, "Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance." Psychological review, vol. 91, no. 3, 1984, p. 328.

[44] I. J. Timm, Dynamisches Konfliktmanagement als Verhaltenssteuerung Intelligenter Agenten (Dynamic Conflict Management as Behavior Control of Intelligent Agents), ser. Dissertationen zur Künstlichen Intelligenz (DISKI). Berlin: Akad. Verl.-Ges. Aka, 2004, no. 283.

[45] W. B. Schaufeli and M. Salanova, "Burnout, Boredom and Engagement in the Workplace," in People at work: An introduction to contemporary work psychology. New York, NY: Wiley, 2014, pp. 293–320.

[46] J. W. Lang and S. Fries, "A Revised 10-Item Version of the Achievement Motives Scale," European Journal of Psychological Assessment, vol. 22, no. 3, 2006, pp. 216–224.

[47] C. Beierlein, A. Kovaleva, C. J. Kemper, and B. Rammstedt, "Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU)," 2012.

[48] J. O. Raynor and I. S. Rubin, "Effects of achievement motivation and future orientation on level of performance." Journal of Personality and Social Psychology, vol. 17, no. 1, 1971, p. 36.

[49] A. Eckhardt, C. Maier, and R. Buettner, "The Influence of Pressure to Perform and Experience on Changing Perceptions and User Performance: A Multi-Method Experimental Analysis," 2012.

[50] R. Buettner, S. Sauer, C. Maier, and A. Eckhardt, "Real-Time Prediction of User Performance Based on Pupillary Assessment via Eye Tracking," AIS Transactions on Human-Computer Interaction, vol. 10, no. 1, 2018, pp. 26–56.

[51] R. Buettner, I. F. Scheuermann, C. Koot, M. Rössle, and I. J. Timm, "Stationarity of a User's Pupil Size Signal as a Precondition of Pupillary-Based Mental Workload Evaluation," in Information Systems and Neuroscience. Springer, 2018, pp. 195–200.

[52] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in Advances in psychology. Elsevier, 1988, vol. 52, pp. 139–183.

[53] F. Man, R. Nygard, and T. Gjesme, "The Achievement Motives Scale (AMS): theoretical basis and results from a first try-out of a Czech form," Scandinavian Journal of Educational Research, vol. 38, no. 3-4, 1994, pp. 209–218.

[54] C. Maslach, W. B. Schaufeli, and M. P. Leiter, "Job Burnout," Annual Review of Psychology, vol. 52, no. 1, Feb. 2001, pp. 397–422.

Challenges in Mitigating Errors in 1oo2D Safety Architecture with COTS Micro-controllers

Amer Kajmakovic*, Konrad Diwold*, Nermin Kajtazovic¶, Robert Zupanc¶

*Pro2Future GmbH & Institute of Technical Informatics, TU-Graz, Graz, AT
E-mail: (amer.kajmakovic, konrad.diwold)@pro2future.com
¶ Siemens AG, Graz, AT, E-mail:(nermin.kajtazovic, robert.zupanc)@siemens.com

***Abstract*—The number of Commercial-Off-The-Shelf (COTS) micro-controllers used in safety applications has increased significantly over the last decade. In contrast to safety-certified micro-controllers, they are produced without integrated protection against memory soft errors and limited in terms of available memory and computation power. However, due to constant optimizations of the memory's physical size and the voltage margins, the probability that external factors, such as magnetic fields or cosmic rays, temporally alter a memory state (and thus cause a soft error) rises. It is crucial to address such errors within safety-critical systems, and consequently a wide range of error mitigation strategies have been proposed. In the context of established brownfield automation systems, redesign and redeployment of new hardware is usually not feasible. Therefore, other approaches can be applied to existing fail-safe architectures to further improve their performance without the need for a partial rework or conceptual changes. This article identifies challenges associated with soft error detection and correction strategies in 1-out-of-2 with diagnostic (1oo2D) safety architecture. Moreover, it investigates mitigation strategies and their deployment challenges through different production phases of the systems (i.e., greenfield) as well as requirements and limitations when working with already existing systems (i.e., brownfield). Among other parameters, the memory usage profile and its effect on the mitigation strategies is explained. A brief overview and evaluation of already available hardware-based strategies along with the evaluation of the most prominent software-based strategies are presented. In addition, a discussion about potential mitigation strategies that rely on the underlying hardware features is outlined. The article demonstrates how to identify and assess trade-offs associated with different strategies to decide on suitable methods to enhance fault tolerance in existing and future automation systems.**

Keywords–soft errors; mixed-criticality; fail-safe; 1oo2D; COTS;

I. Introduction

This article extends the contribution "Challenges in Mitigating Soft Errors in Safety-critical Systems with COTS Microprocessors" of PESARO 2020 [1]. The contribution investigated challenges associated with software-based soft error detection and correction strategies, along with a short overview of currently applicable software-based mitigation strategies. Here, the evaluation is extended to include available hardware-based strategies and different phases in the development process of 1oo2D safety architectures. Furthermore, new ideas and approaches are presented utilizing existing features within 1oo2D architectures to avoid physical intervention on the system.

Given their ever-decreasing packaging size, semiconductors are increasingly susceptible to external influences such as alpha particles, cosmic rays, or magnetic fields [2]. Figure 1 shows the correlation of semiconductor technology/fabrication node size (nm) and their respective error rates (Soft Error Rate (SER) and Hard Error Rate (HER)). It is evident that the SER increases with decreasing node size, while the HER remains constant [3]. To counter the increasing number of soft errors, families of highly reliable safety-certified Micro-Controller Units (MCUs), with special integrated measures against soft errors, have been developed. The intended field of application of such micro-controllers is safety-critical applications where fault-tolerance is required.

Nevertheless, given their low cost and good performance, Commercial-Off-The-Shelf (COTS) micro-controllers are increasingly used in safety applications [4]. In contrast to safety-certified micro-controllers, they are not produced with integrated protection against soft errors. As a consequence, recent research proactively deals with environmentally induced soft errors by developing new methods for error detection, mitigation, and data recovery [5].

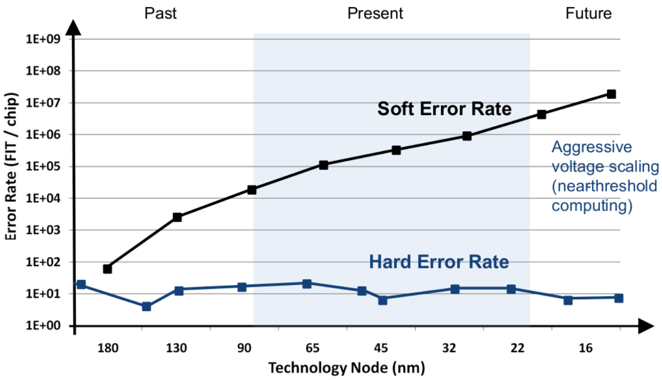


Figure 1. Correlation of error rate and technology/fabrication nodes

The importance of detecting and resolving soft errors is reflected by the numerous reports on soft error related problems within safety-critical applications. These reports originate from a wide range of industries, such as the automotive industry, space industry, and the medical industry. Duncan and Roche's analysis of semiconductor reliability in the context of autonomous driving [6] is devastating. They conclude a (soft error induced) failure rate of 1 part per million per year. Given that a single-car implements approximately 8,000 semiconductors, the likelihood of a car exhibiting semiconductor-induced errors within its lifespan (of 15 years) is around 12%. While the results of such failures are unclear during the operation of a car, semiconductor-based soft errors can be resolved (fairly easily) by restarting the affected component. However, not all safety-critical systems provide the luxury of resolving an error by "turning it off and on again". Consider, for example, safety-critical nuclear power plant equipment: restarting a device in the event of a soft error is not an option and could lead to catastrophic fatalities.

When designing new safety-critical applications or enhancing existing systems with fixed underlying architectures and hardware, a wide range of methods is available. These methods target different stages within a system's development and are associated with trade-offs regarding required resources, costs, and complexity. To choose an appropriate strategy therefore requires a clear understanding of the available methods as well as their prerequisites. This article aims to demonstrate the variety of existing mechanisms for soft error detection and correction, by reviewing and outlining available methods. In addition, the article demonstrates how an appropriate methodology can be chosen, depending on the development stage of the application along with challenges that come with selecting the right strategy.

While architects are 'relatively' free when designing a new system from scratch, their options narrow down when they are enhancing an existing system, as the chosen methods must complement the existing system. Given the longevity of existing safety automation solutions, this article demonstrates an approach to improve/enhance existing fail-safe solutions. This is done utilizing an exemplary system with a 1oo2D safety-architecture and allows to demonstrate the impact and prerequisites of various safety strategies on the system's performance and design as well as their effects on non-functional requirements, such as reliability, safety, and availability. The discussed approaches range from enhancing the existing hardware solutions with additional software-based correction schemes to the utilization of additional hardware, resulting in novel hybrid approaches. These innovative approaches allow enhancement of non-functional properties such as availability, maintainability, and most importantly safety in existing safety architectures.

The remainder of the paper is organized as follows: Section II presents an overview of the mitigation strategies through the production phases. In Section II-E a screening of the market-available micro-controllers with mitigation strategies is presented. Section III describes 1oo2D safety architecture with a focus on its memory architecture. Section IV defines the challenges and requirements for soft error software-based mitigation strategies in safety-critical applications. Section V shows an evaluation of the mitigation strategies along with new mitigation ideas. In the last section, a summary and future work are presented.

II. Mitigating soft errors

While soft errors constitute the majority of memory errors, they can be prevented and/or corrected. To prevent soft errors, memories require resilience and/or fault-tolerance. Fault-tolerance denotes a system's ability to handle faults in individual hardware or software components, power failures, or other forms of unexpected problems, while still meeting the system specification [7]. There are different approaches/strategies to achieve fault tolerance. These approaches can be grouped into different levels regarding the stage in the development process they are utilized in as well as their underlying nature.

The most intuitive categorization can be made based on the different stages of a system's development. Protection and mitigation strategies can be designed and applied within the design and production processes of single components (i.e., memories) themselves. During system design, mitigation strategies can be actively integrated into the system by, for example, choosing appropriate components and system architectures (such as redundant architectures). If a system's architectural level has already been fixed (during or before the deployment stage), only software-based approaches can be used to enhance fault-tolerance on a system level (e.g., via additional features that will additionally secure a system). During system design, fault-tolerance mechanisms on a hardware-level (e.g., by hardening components and architecture) can also be utilized. Mitigation strategies thus either fall into the Hardware-based (HW) or Software-based (SW) classes. They are not mutually exclusive, meaning that a system might implement a set of different mitigation strategies to achieve required fault tolerance. Another categorization concerns whether or not an approach utilizes redundancy. Within a system, redundancy can occur on different levels: Hardware, Software, Information, and Timing, which are explained in more detail below. Figure 2 gives examples for mitigation strategies and their categorization.

Stages	Type	HW/SW	Examples
Component level	Early design stage	HW	Hardening of the component cases
Architecture level	Hardware redundancy	HW	2oo3 architecture
	Informationa l redundancy	HW	Memory with HW parity bit protection
System level	Software redundancy	SW	Parallel execution of redundant function
	Informationa l redundancy	SW	SW ECC code
	Timing redundancy	SW	Repetition of the same function execution

Figure 2. Categorization of the mitigation strategies

In the following subsections, state of the art mitigation strategies are outlined according to the system development levels they fall into, starting with the system level.

A. System level

Protection on the system level is applied when the hardware of a system is present, including internal design and system architecture. At this level, additional fault-tolerance can only be achieved via software-based approaches (as hardware and system architecture are fixed). Methods applicable to this phase can also be used to enhance existing (brownfield) automation systems that are already deployed and do not allow for hardware changes.

Software-driven fault-tolerant techniques are based on redundancy, which is applied to procedures, processes, data, or the whole execution code. The most common type of **software redundancy** in embedded systems is the multiplication of data. A simple way of achieving multiplication is to transform (e.g., with the Hamming distance of 4 or a simple inverse function) and store a copy of a variable in a different memory area. Comparison of the two versions of the variable enables the system to detect, mitigate, or recover corrupted data.

The main disadvantage of software redundancy is associated with memory consumption overhead, as the multiplication of data, code, and/or processes requires additional memory, which is usually very limited in embedded systems. Additionally, it can lead to a significant increase in code execution time [2], [5]. The other two types of redundancy which can be realized in the software itself are Informational and Timing redundancy.

Informational redundancy assumes the addition of supplementary information to the data to verify the soundness of the information. Usually, this additional information is in the form of codes, which are computed based on the data itself. Those codes (so-called Error Detection And Correction codes (EDAC)) were initially introduced in the context of data recovery in communication [7], but nowadays they are widely used in memories [8]. The family of EDAC codes is growing constantly. The most popular EDAC codes are: Parity Codes (error detection without recovery) [9], Hamming Codes (2-bit detection and 1-bit recovery) [9], Reed-Solomon and Bose-Chaudhuri-Hocquengham Codes (for multiple bits error masking) [8]. Some research has considered the implementation of other EDAC codes used in communication such as LDPC codes [10], RS codes, Turbo codes [11]. EDAC codes can be presented with the designation (n, k), denoting a block code that takes a k-bit data word and maps it to an n-bit codeword as shown in Figure 3.

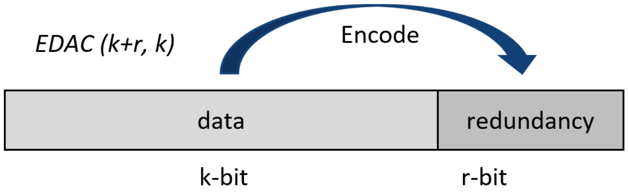


Figure 3. Representation of EDAC codes

EDAC codes have two main properties that need to be considered: speed and quality. Speed is defined as the time required to encode/decode EDAC codes and this time extends the overall memory access time. Quality denotes the number of faulty bits a specific code can detect and correct. Naturally, there is a trade-off between quality and speed. For higher quality, more complex EDAC codes are required, which allow for correction of multiple bit-flips. In this case, both code magnitude as well as computing demand increase due to these adaptations. Faster and less memory expensive correction schemes on the other hand are limited in terms of the number of bits they can correct.

Based on EDAC codes, a new method called **scrubbing** was developed. The idea behind scrubbing is to periodically re-write data in its original location to eliminate soft errors if they are correctable through EDAC [12] or copying of original data [13]. With this approach, an accumulation of soft errors inside one region of memory can be avoided.

Timing redundancy has been recently investigated and concerns a re-computation or retransmission of data at least twice. The results are then compared with previously stored copies [7]. This type of redundancy helps to distinguish between transient and permanent errors. If the fault is still present after repeating a test several times, then it is likely that the error is permanent.

B. Architecture level

HW-based Information redundancy: Software-based information redundancy raises the question of usability, as high-quality SW EDAC codes exhibit a trade-off and lead to a decrease of available memory as well as to an increase of required computation time, access time, and complexity of the overall system. To overcome these drawbacks, EDAC-related computations (encoding and decoding) can be outsourced on a special-purpose chip, which can be installed between memory and CPU in order to apply for on-the-fly informational redundancy. Most modern EDAC codes for memories are implemented via additional hardware [14]. EDAC addresses the perspective of system availability for safety, since the system will continue to run unabated in the presence of single-bit errors. However, EDAC adds significant cost to the memory portion of the device and slows down the CPU due to the added SRAM access time, which is required to make corrections on the fly. SRAM on a device constitutes about 1/3 of the hardware costs, and with additional HW-based EDAC this further increases by approximately 30%, resulting in a total price increase of around 40% [15]. To avoid an increase in chip size and hardware redesigns, software-based EDAC codes (explained in the previous chapter) have been proposed [16], [17]. In the past, HW EDAC codes were only available in the expensive safety-certified MCUs, but today conventional micro-controllers also possess HW-based EDAC protection. The flash memory, where operating code is stored, is usually protected with a Hamming code while parity bits protect selected parts of the SRAM [18], [19]).

A parity circuitry sets the parity bits when an SRAM word location is written and verifies that there are no single-bit errors in the word when it is read back. This is done within the read/write cycles, so no CPU overhead is involved. When the parity circuitry identifies an error, a high priority CPU interrupt is generated. In semiconductor devices, this detection mechanism is simple and relatively inexpensive to implement. Parity addresses the safe-state perspective for safety. As described earlier in Section I, virtually all SRAM failures in-system are likely to be single bit per word failures. This applies to both physical defect mechanisms as well as soft errors. Additional coverage can be provided by protecting the memory address bits with parity.

Hardware Redundancy: On a system level, fault tolerance can be achieved via hardware redundancy. Safety-critical systems often adopt an N-modular (where $N > 2$) architecture, where the components exist in certain redundancy N and perform the same computations in parallel. The correct result is established based on majority voting. If one of the modules fails, the majority voter masks the fault by identifying the result of the remaining fault-free modules [7]. N-modular systems can yield a higher Safety Integrity Level (SIL), as they provide inherent fault tolerance and consequently, a low failure rate. SIL is a quality indicator for systems that fulfill safety requirements in accordance with the IEC61508 standard. Many safety systems use simple architectures such as 1oo1D (1-out-of-1 with diagnostics) and 1oo2D (1-out-of-2 with diagnostics) [20]. In some cases, a diagnostic system is realized with an additional watchdog (i.e., challenge-response architecture) [21] or with an additional CPU like the lockstep architecture.

Lockstep systems are fault-tolerant computer systems that run the same set of operations at the same time in parallel [22]. The redundancy (duplication) allows error detection in the system as well as in the memories. The stored values in memory are compared to determine if there has been a fault. The term ”lockstep” originates from army terminology, where it refers to synchronized walking, in which the marchers walk as closely to each other as physically possible.

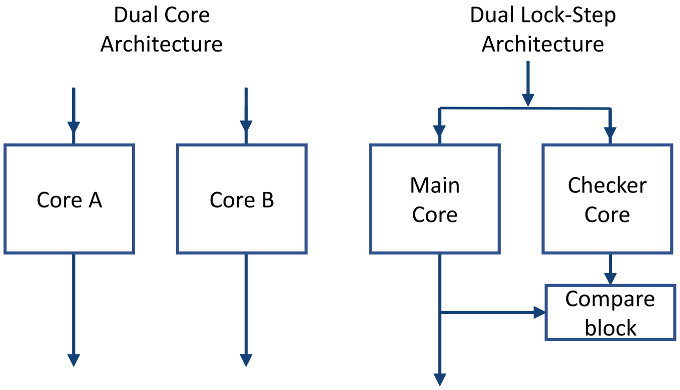


Figure 4. Dual Core and Lockstep architectures

These architectures are also known as a fail-safe, meaning that given a failure, the system inherently responds in a way that will cause no or only minimal harm to equipment, environment, and people. The main advantage of such architectures is the good balance between functional safety (i.e., achieving high safety integrity) and development process costs.

A shortcoming of hardware redundancy is its requirement for additional hardware. In the context of memory, it will increase cost, weight, size, power consumption, and thus impacts design and testing. Moreover, additional hardware needs to be budgeted for from the first stage of the chip design. It is therefore almost impossible to upgrade existing systems with additional hardware without degrading their performance, which limits the application of these methods in the context of brownfield applications.

C. Component level

Environments with high ionizing radiation (e.g., outer space, nuclear power plants, etc.) present special design challenges for integrated circuits, as the likelihood that particles cause an upset in the electronics (i.e., memory) is very high. Dealing with the consequences requires very reliable electronic components with sophisticated measures that can detect and correct errors. The first step in overcoming errors is to prevent them from happening, i.e., to stop particles on their way to the sensitive parts of the electronic circuits. This type of protection is achieved during the early stage designs, where different techniques and approaches are used to prevent errors. If these techniques can successfully protect electronics, in later phases they do not need additional detection and correction algorithms.

Shielding constitutes one of the first approaches that increase the resilience of components against radiation. Shielding is applied during the production phase, where a specific particle-resistant layer is deployed over the component’s package. The layer reduces exposure of the bare component/device and prevents environmental particles from influencing underlying layers of the package. Figure 5 depicts the penetration ability of various types of particles. As shown in the image, neutrons are capable of travelling further through different types of material than other particles, making it challenging for designers to find adequate materials for shielding.

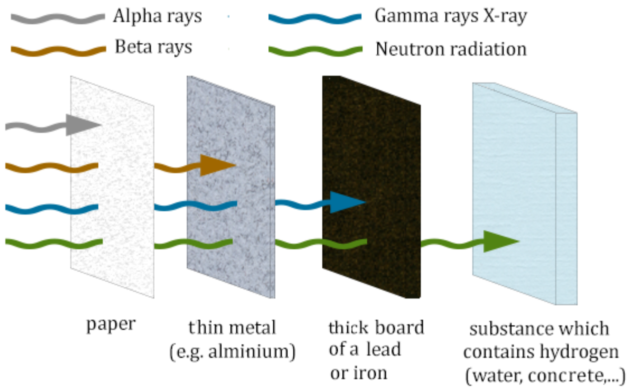


Figure 5. Penetration ability of radiation particles [23]

Radiation hardening is an approach where designers of electronic circuits use various physical means, such as insulating substrates, bipolar integrated circuits, or radiation-tolerant SRAM to harden the electronic system against the effects of radiation particles [24]. Hardened chips are often manufactured on insulating substrates instead of the usual semiconductor wafers (where energy from radiation can easily change the state of the material). Silicon on insulator (SOI) [25] and Silicon On Sapphire (SOS) [26] are commonly used. While hardening guarantees fewer errors to be caused by radiation, it requires special designs and techniques that increase the overall costs of the design and production process. Resistance to electrical charges can also be achieved by using specific structures and materials for critical points in the component (e.g., strengthening the gate of the transistors). One of these structures is the Dual Interlocked Storage Cell (DICE). In this technique, a transistor structure has redundant storage nodes and restores the original cell state when an error is introduced in a single node [27].

Other types of memories that are not based on standard semiconductors but on different underlying concepts can be found. The most promising concept is Phase-Change memory (PCM), which constitutes a new type of memory that is achieving good results against particle radiation. PCM utilizes a Germanium Antimony Tellurium $Ge_2Sb_2Te_5$ (GST) alloy and takes advantage of rapid heat-controlled changes in the material’s physical property of amorphous and crystalline states [28]. These states, which correspond to logic 0 and 1, are electrically differentiated by high resistance in the amorphous state (logic 0) and low resistance in the crystalline state (logic 1). One cell of the PCM is shown in Figure 6. PCM, which reads and writes at low voltage, offers several substantial advantages over flash and other embedded memory technologies: PCM is faster than standard flash memories, and logical gates within PCM can be scaled down further than the NOR and NAND gates used in flash memories. PCM also showed good protection against bit-flips induced by highly energized particles hitting the memory. Even though phase change material is immune to high-energy particles, PCM memory still suffers soft errors. For example, in PCM chips, up to 40% of the entire area consists of CMOS circuits [29]. As

PCM is still in development, only a few types of this memory are available on the market [30]. Similar to radiation hardening, PCM memory is used in the early design stage of the MCU, and thus, it does not have additional effects on the MCUs' performances.

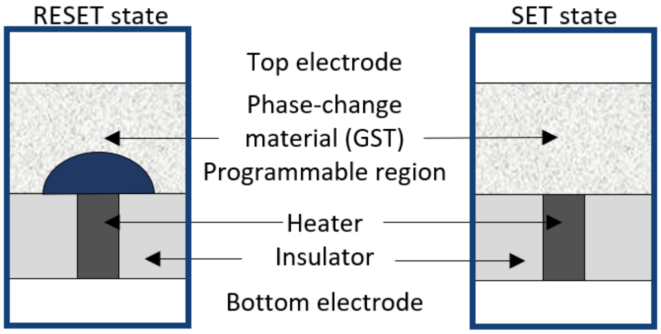


Figure 6. Phase changing memory - Reset and Set states [31].

Although techniques used on a component's level have shown very effective against soft errors, they always require additional or special materials, which significantly increases the cost of design and production.

D. Multi-phase

Some approaches to the production phase (i.e., component level) can also be used in later stages. For example, shielding can be applied after the entire system is developed, e.g., by installing a radiation-resistance shield over the system itself. Combining the best features of different protection schemes, to cover their weakness, constitutes a good way to create system tolerance for all kinds of failures. Mayuga et al. [12] combined different kinds of techniques to overcome failures in memory. Their approach uses EDAC codes to recover words with a single faulty bit, memory relocation for a word with more than one faulty bit, and a scrubbing method to avoid the accumulation of faulty bits. A hybrid approach seems very suitable in the context of mixed-criticality, as it allows to further customize the overall protection scheme, leading to a protection scheme with an even further reduced overhead in comparison to a scheme that is based on single redundancies.

E. Available solutions for safety-critical systems

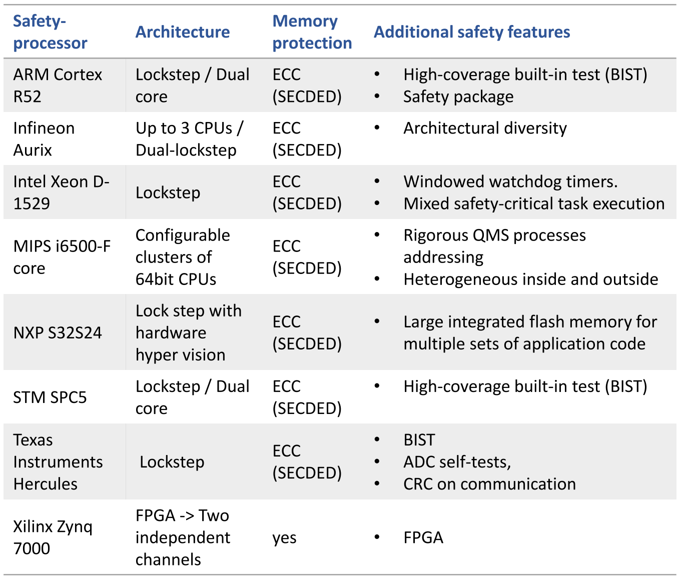
When designing a safety-critical system from scratch, it is recommended to proactively consider soft errors through all production phases, in order to satisfy the safety requirements of the resulting system. As shown in the last section, designing a system from scratch allows us to utilize hardware solutions to mitigate or overcome errors, e.g., by choosing an appropriate architecture or components that already have integrated safety measures against soft errors in memories.

Manufacturers have developed special-purpose safety-certified micro-controllers that are highly reliable and contain additional features to overcome safety issues, including soft errors in memories. The design of these micro-controllers demands more time and effort, thus their development is far more costly than regular COTS micro-controllers. The main catalyst for these recent developments is the automotive industry. With high demands for functional safety in Autonomous (AV) and Semi-Autonomous Vehicles (SAV), the development of safety-critical micro-controllers has rapidly increased. Functional safety is required in almost every part of AV and SAVs, including all sensors, processing, and control units. Some MCU developers like STMicroelectronics are offering a wide portfolio of MCUs specialized for automotive applications. The latest achievement in safety from STM are the controllers from the Stellar series, a high-performance 32-bit automotive microcontroller family, which is based on the ARM R52 multi-core. It features an innovative embedded Phase Change Memory (ePCM) and built-in 28nm Fully Depleted Silicon On Insulator (FD-SOI) technology [32]. The combination of ECC and this memory can provide sufficient protection from soft errors in these processors.

In the context of automotive use-cases, the probability of a particle hitting the memory and flipping a bit is low, but the impact of a bit flip can be devastating. In space and the nuclear industry, besides devastating impact, the probability of a particle altering the memory is very high. Therefore, the need for radiation-resistant electronics is higher than in any other domain. This can be achieved e.g., via HARDSIL® a special technology that immunizes semiconductor devices against high temperatures or radiation-induced stress without the need for special design techniques (RHBD) or expensive specialized semiconductor processes (RHBP). HARDSIL® can enhance a broad range of semiconductor devices. It is a fully designed agnostic approach, where any standard manufacturing equipment and process geometry can be used with no resulting negative impact on performance, power consumption, or yield. Simulations have shown the ability to scale down to the most sophisticated leading-edge technologies like Fin-FET implementations [33]. One type of MCU that employs HARDSIL technology is the VA108X0 [34] micro-controller from Vorago technologies, based on the ARM®Cortex®-M0 processor with a radiation tolerant case.

Another example of highly reliable MCU is the TMS570 series from Texas Instruments' line Hercules. This safety micro-controller is targeted for safety applications, through hardware-based fault correction/detection features in the form of dual cores that can run in lockstep. Moreover, it has automated self-testing of memory and logic, peripheral redundancy, monitor/checker cores, and full path ECC. The full path ECC means that all memories in the MCU are protected with ECC (Flash, Data Flash for EEPROM, SRAM). This type of hardware-based ECC is performed by the CPUs and can correct single-bit errors and detect double-bit errors (SECDED). The ECC is evaluated in parallel to application processing, so there is no impact on latency or performance. The same integrated safety protection can be found in NXP's Kinetis Kx line. Their Flash and RAM memories are also protected with ECC codes (SECDEC).

The examples and Figure 7 outline a selection of MCUs specialized for safety-critical applications. These MCUs are certified according to automotive (ISO 26262[35]) and industrial (EC 61508[36]) safety standards. As a result, they are significantly more expensive than standard COTS MCUs. This cost difference may lead engineers and designers of safety systems to look at cheaper solutions based on COTS MCUs. As outlined above, some COTS MCUs already integrate simple hardware memory protection such as parity bits. The computation of these simple protection schemes requires less resources than more complex EDAC (SECDED) codes. Adding them to a controller does not significantly increase the overall costs. The



Safety-processor	Architecture	Memory protection	Additional safety features
ARM Cortex R52	Lockstep / Dual core	ECC (SECDED)	• High-coverage built-in test (BIST) • Safety package
Infineon Aurix	Up to 3 CPUs / Dual-lockstep	ECC (SECDED)	• Architectural diversity
Intel Xeon D-1529	Lockstep	ECC (SECDED)	• Windowed watchdog timers. • Mixed safety-critical task execution
MIPS i6500-F core	Configurable clusters of 64bit CPUs	ECC (SECDED)	• Rigorous QMS processes addressing • Heterogeneous inside and outside
NXP S32S24	Lock step with hardware hyper vision	ECC (SECDED)	• Large integrated flash memory for multiple sets of application code
STM SPC5	Lockstep / Dual core	ECC (SECDED)	• High-coverage built-in test (BIST)
Texas Instruments Hercules	Lockstep	ECC (SECDED)	• BIST • ADC self-tests, • CRC on communication
Xilinx Zynq 7000	FPGA -> Two independent channels	yes	• FPGA

Figure 7. Safety-certified (IEC61508, ISO26262) micro-controllers

parity bit is a prime example of a simple protection scheme, however, it comes with drawbacks as it can only detect odd-numbered bit errors (single, triple, etc.) in the protected word. In addition, using parity bits only allows the detection of the error, and and without a proper safety architecture, memory recovery is practically impossible.

For example, the STM32L4 series and some MCUs of the STM32Fx series from ST Microelectronics have parity protection for 25% of their memory in addition to EDAC codes for flash memory. When an error occurs in protected memory, an interrupt with high priority is activated on the CPU side. In this way the error is detected, but there is no way to recover it. Advanced MCUs have additional ECC protection for SRAM memories. For example, in the case of Cortex R4 based CPUs, the EDAC encoding/decoding is done by the CPU (in-built), whereas in the case of Cortex-M3 and ARM7TDMI-based CPUs, the ECC encoding/decoding is done by the RAM wrapper. The main advantage of having the encoding/decoding within the CPU is to speed up memory access by removing the time-consuming SECDED block. SECDED is based on the Flash/RAM technology design of the controller and is adapted accordingly. Some designs have two SECDED modules that operate in parallel. The results are then compared and accepted only if both are the same [37].

III. Prevailing safety architectures

Safety-critical systems often adopt an N-modular (where $N > 2$) architecture. The components exist in certain redundancy and perform the same computations in parallel. The correct result is established based on majority voting. If one of the modules fails, the majority voter masks the fault by identifying the result of the remaining fault-free modules [7]. Although N-modular systems can achieve a higher SIL level, as they provide inherent fault tolerance and consequently a low failure rate, many safety systems use simple architectures such as 1oo1D and 1oo2D [20]. The main advantage is that they have a good balance between functional safety (i.e., achieving a high safety level) and development process costs.

In 1oo2D architectures, all hardware including sensor inputs is independently implemented twice. This leads to a multi-core architecture similar to the one described in [38]. The output of these parallel lines is checked and selected by a voter [39]. For a safety system, it is quite often not important if the final result (chosen by the voter) is correct, as long as it is safe. In case the two outputs differ, the result leading to a safe and non-critical state is preferred and opted for by the voter.

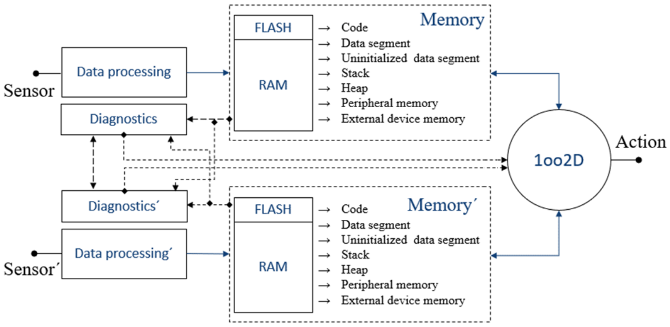


Figure 8. Memory model in 1oo2D architecture

For memory, a 1oo2D architecture provides independent memories for each parallel line of the computing system. Two independent parallel memories ensure system hardware and software redundancy. This means that besides memory-specific data which is required for synchronization, identical data can be found on both memories (Figure 8 depicts the memory model in a 1oo2D architecture). Data in mixed-critical memories can be categorized into safety-relevant and safety non-relevant data. The different criticality levels of data in combination with duplicated memory in the 1oo2D architecture are shown in Figure 9.

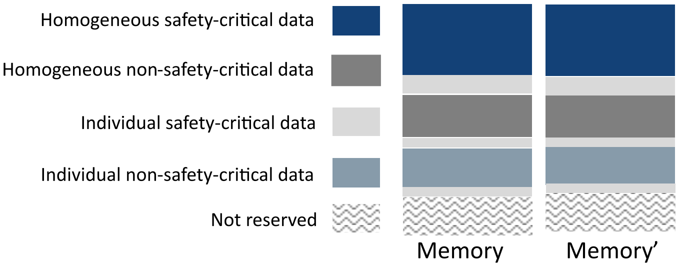

Homogeneous safety-critical data
Homogeneous non-safety-critical data
Individual safety-critical data
Individual non-safety-critical data
Not reserved
Memory
Memory'

Figure 9. Example of the mixed critical memory in the 1oo2D safety architecture

All regions are equally exposed to faults, however, different forms of protection can be applied to different regions. Experts advise that protection should be implemented in the form of periodical test runs over data. As a guide, we refer the reader to the Safety manual [40] provided by STMicroelectronics for their micro-controllers. To enhance the coverage of hard errors on SRAM, detection tests like Galloping [41] or March classes [42] have been proposed.

For soft errors, STMicroelectronics advises redundancy to be implemented for all safety-relevant variables. Typical solutions provide a copy of original data on the same memory chip or on an additional (redundant) chip. The copied data is periodically compared to the original, in order to detect the presence of errors [43]. If an error is detected, it is not clear which memory (or part of the memory) is affected. Hence, such a solution leads to detection but not a correction of the soft error and will result in the system transitioning into a safe-state.

As we have seen in Figure 1, the number of soft errors shows a negative correlation with the size of the underlying transistors, leading to a rapid increase of soft errors. In the context of automation, this means that systems are more likely to go into safe states that can disrupt or stop the automation process. These unwanted halts affect the availability of the system [44]. A solution for overcoming this problem is to add mechanisms on top of the existing architecture, which allow for the recovery of faulty data and to extend the on-line time of the system. Recovery mechanisms in this context are usually ECC based. As outlined before, ECC is mostly hardware-based and requires additional time and additional hardware for computing. It is not feasible to extend existing brownfield automation systems with additional hardware, as this would lead to the need for a complete redesign of the system. Another option is to apply software-based ECC approaches, which are complex and expensive in terms of computation.

Given that 1oo2D already provides the possibility to detect memory errors, the question arises how existing architectures (i.e., 1oo2D) can be combined with other approaches that allow correction of detected soft errors and exhibit very little overhead. In addition, these methods should be flexible in terms of their configuration, to enable their application in the aforementioned mixed-critical scenarios where safety-critical data requires more detailed monitoring.

From previous discussions it seems obvious that a solution enables better memory error detection and correction strategies for existing automation products must take the best of two worlds, i.e., utilizing the properties of the underlying architecture to the fullest extent and combining them with flexible software-based soft error correction methods which show little overhead and can be adjusted in terms of mixed-criticality of the prevailing system.

IV. Challenges in Mitigating Soft Errors

To overcome soft errors and consequently lower their impact on the non-functional properties of a system, various methods for error detection, correction, and mitigation were introduced. As already stated in the previous section, the available methods can be divided into hardware- and software-based correction mechanisms. Hardware-based mechanisms provide error detection and correction on an architectural level and use specific hardware. Hardware approaches are not applicable in the brownfield, i.e., existing devices or systems, and usually involve redesign and redeployment. For brownfield systems or devices, software solutions fit better because they can be implemented with a simple update or software patch and consequently minimize costs. Software-based correction mechanisms operate on the memory itself without altering the underlying hardware or architecture. Depending on the application, adequate correction quality is required. Quality denotes the fault magnitude that the strategy is capable of detecting, mitigating, and/or recovering. Given that there is no such thing as a free lunch, soft error strategies require additional execution time and/or memory space, and therefore affect processor run-time and can cause increased memory overhead. On the other hand, hardware-based strategies are more reliable, more powerful, and faster when it comes to computing EDAC codes.

These observations lead to a general trade-off problem for the design and deployment of error detection and correction, as it is always required to balance the quality of detection (required by the underlying application) and the resources required to implement appropriate correction and detection strategies. Higher quality error correction requires more computation time, more memory capacity, and sometimes additional hardware. Depending on the target system, this might lead to a violation of the system's requirements in terms of cost, available memory space, or computation time of the system's applications. In the following, the system requirements are outlined in more detail.

1) Run-time performance: The development of methods, which provide sufficient error coverage, while keeping the impact on a system's run-time or memory overhead minimal, is particularly important in the context of safety-critical systems. This is due to the fact that such systems have very strict timing requirements (i.e., norms in the field define specific timing limits, such as Fault Tolerant Time Interval (FTTI) (see Figure 10) in ISO26262 or Process Safety Time (PST) in the IEC61508 standard). The FTTI constitutes the time-span between a fault and the hazard which results from it [36], [35].

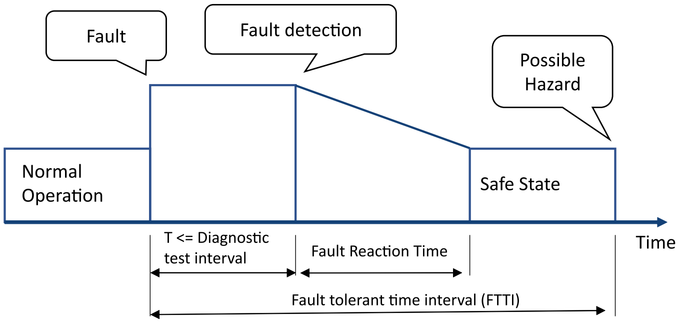
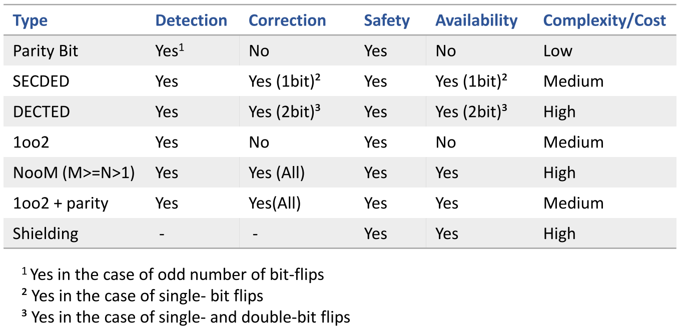


Figure 10. Fault reaction time and Fault Tolerant Time Interval (FTTI) [35]

Faults must be detected and corrected within this interval. If a correction is not possible, the system must be guaranteed to reach a safe state within the FTTI. Therefore, the run-time performance of correction strategies plays a crucial role in the context of safety-critical systems, as its application must not lead to a violation of these FTTI requirements. For example, when using software calculated EDAC codes, the computation time required to calculate redundant bits needs to be evaluated and taken into consideration. If additional hardware is calculating redundant bits, it will increase memory access. This time will probably not significantly affect overall run time, but engineers need to be aware of it [15].

2) Memory consumption: Many software-based strategies require additional memory space for their implementation, which is used to store copies of data or code, or additional information required by the method, such as Parity bits or EDAC. Compared to a similar software solution, EDAC codes exhibit the smallest overhead. The ratio between additional bits required for protection and protected bits is always less than one in EDAC, whereas this is not the case for full redundancy. While in most cases EDAC codes can have a large memory footprint, parity bits constitute their most lightweight form. They allow monitoring of the consistency of a memory region with a defined length based on a single bit, which denotes whether the number of one-bits in the region is odd or even. Decreasing the size of the protected region can lead to increased memory overhead. To give an example: the protection of a 32-bit word via Hamming code will result in a 3.15% memory overhead. One-bit recovery of a 32-bit word, using Hamming code, would require an additional 7 bits and result in a memory overhead of 22%. The EDAC calculation always requires additional hardware components that will do the calculations and store the calculated bits. Different redundant architectures also require additional components, or entire multiplied systems as is the case with 1oo2 or 2oo3 safety architectures.

3) Mitigation quality: The quality of a strategy is defined by its capability to detect and correct (i.e., recover from) faulty bits. A system's detection and correction capabilities are reflected in the number of faulty bits that can be detected and corrected. The simplest EDAC code (Parity) can detect all odd-numbered bit flips but does not provide recovering capabilities. A 2oo3 system can detect and correct all bit flips, but its complexity and consequently costs are much higher. A short overview of the quality for some mitigation strategies is given in Figure 11



Type	Detection	Correction	Safety	Availability	Complexity/Cost
Parity Bit	Yes[1]	No	Yes	No	Low
SECDED	Yes	Yes (1bit)[2]	Yes	Yes (1bit)[2]	Medium
DECTED	Yes	Yes (2bit)[3]	Yes	Yes (2bit)[3]	High
1oo2	Yes	No	Yes	No	Medium
NooM (M>=N>1)	Yes	Yes (All)	Yes	Yes	High
1oo2 + parity	Yes	Yes(All)	Yes	Yes	Medium
Shielding	-	-	Yes	Yes	High

[1] Yes in the case of odd number of bit-flips
[2] Yes in the case of single- bit flips
[3] Yes in the case of single- and double-bit flips

Figure 11. Mitigation strategies and quality parameters

In fail-safe systems, detection of an error is usually reflected with the safety feature because detection is enough to trigger activation of the safe state, which prevents further safety issues. Between error detection and safe-state activation, the system has a defined allowed time for recovery. If recovery is not possible for any reason, the system will transition into the safe state and its availability will be affected.

4) Mixed criticality: Safety-critical applications usually exhibit different levels of criticality in terms of their underlying data. While a fraction of data is system critical (i.e., if affected by an error the consequences can be catastrophic), errors affecting non-critical data will not impact the safety of operation. This phenomenon is known as mixed-criticality [45]. Incorporating mixed-criticality into the design of mitigation strategies, by devising and applying different detection and correction strategies on memory areas holding data of different levels of criticality, allows further improvement of a system's availability while guaranteeing a correct treatment of system-critical events [45]. While adequate protection needs to be provided for the whole system, safety-critical data requires stronger protection. Several recent studies have investigated mixed-critically in memories, with a focus on data delivery and prioritization according to data criticality [46].

Taking mixed-criticality into account when designing memory detection and correction strategies allows the reliability and safety of the underlying system to be enhanced, as such strategies aim to increase the protection of safety-critical memory parts. By defining different parts of memory to have different criticality, the overhead of correction strategies can be reduced, in contrast to applying rigid correction/detection strategies to the entire memory. In addition, incorporating mixed-criticality can increase a system's availability, as faults in non-system critical memory areas will not necessarily lead to a halt of the system. Figure 9 shows the example of mixed critical memory in the 1oo2D safety architecture.

5) Frequency of access: One interesting phenomenon that can be discussed in the context of protecting memories is access frequency. Two classes of memory access can be distinguished here: low-frequency and high-frequency memory access [47]. Memories with high frequency are more general purpose and can be updated several times per execution cycle.

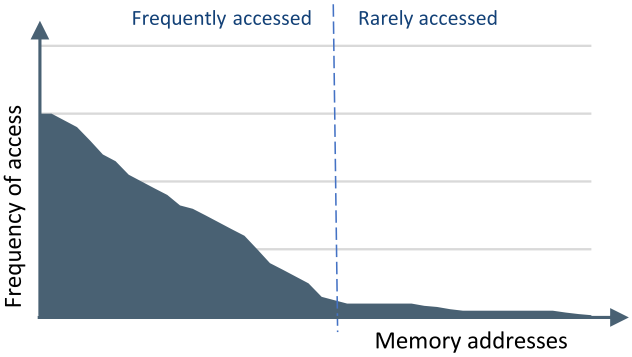


Figure 12. Sketch of potential memory usage profile

The parts of the memories that have lower access frequency usually include on-demand or periodically accessed data, with large time intervals between consecutive accesses. Memories used on-demand could, for instance, store the address of a function that takes the system to the safe-state. As safe-state activation does not happen often, the function will remain unused for long periods of time and thus will not be tested often. Nevertheless, it must always be available. The accumulation of soft errors on these resources can be of high relevance, for example, when the system needs to comply with specific normative requirements (e.g., SIL3 according to the IEC61508 standard [36]). Figure 12 depicts an exemplary memory usage profile. To obtain a realistic memory usage profile, a safety-critical device must be analyzed, as memory usage depends on the applications.

To give an example let us imagine a system with hardware-integrated parity bit protection. Parity bit protection can only detect odd bit flips (single, triple, etc.) without correction, and detection is only triggered when the protected part of memory is accessed. In the context of rare access, the possibility exists that this part of memory experiences more than two separate bit flips between two accesses (protection activation). This can lead to errors going undetected at the next access and can lead to an unsafe-state of the system. The explained scenarios and the effect of accumulation are shown in Figure 13. In sequence (a), an error will be detected, because the parity bit will not respond to the data, while in sequence (b), the test will not detect an error in data because the calculated parity bit responds to the data. This second phenomenon is called the accumulation of error.

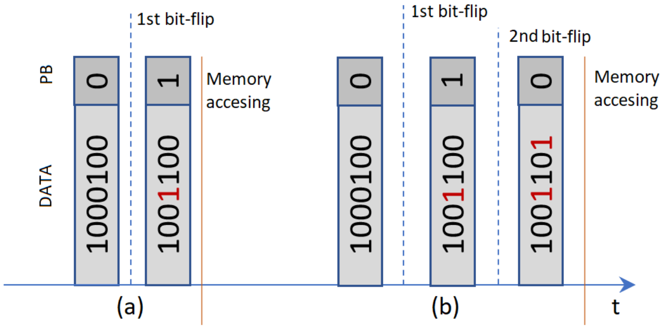


Figure 13. Sequence of the events: (a) when error will be detected , (b) when error will go undetected

6) Memory organization: Due to environmental changes, occurrences of soft memory errors are not continuous, and the chance of a cell being hit by an error is randomly distributed. Errors can appear at any time and in any type of memory or memory part, which can aggravate protection and detection mechanisms as they are type-dependent. One can distinguish between two types of memory in embedded systems: non-volatile and volatile memory. Non-volatile memory sustains stored information during a loss of power (e.g., flash memory), while volatile memory requires constant power to retain stored information (e.g., SRAM) [48].

Embedded memories exhibit various regions: program memory, data memory, registers, and I/O ports [49]. From a software point of view, the memory layout of C/C++ programs consists of the different sections that are saved in different memory regions. Typical memory representations of C/C++ programs consist of a code segment, data segment, uninitialized data segment (bss), stack, and heap. All of this can impact the design of correction/mitigation mechanisms.

7) Availability vs Safety: Safe-state activation often leads to a functional degradation of many system components, and as it often results in a system halt it is associated with high costs. It decreases the availability of the system to ensure the safety of the system and its environment.

Especially in production lines, where every minute without service is associated with high costs, a system's availability is of utmost importance. However, a highly available system is costly, because it demands complex redundant architectures. Therefore, a trade-off between safety and availability exists, that needs to be optimized. One way to increase availability while keeping functional safety on the demanded level is to postpone or avoid the unnecessary activation of safe-states. In [44] the concept of Predictive Fail-safe was proposed, which aims to increase a system's availability by applying data analytics on safety-relevant data to predict and prevent future failures.

8) Usability: Soft errors have been a focus of research for the past 60 years. Although many approaches have been introduced and tested with good results, only a few have found their way into real-world applications. This is due to the associated required resources (e.g., computation power and time or memory consumption), which limits their applicability. The strategies outlined in Section II-E are the only ones that are currently applicable given the performance capabilities of available micro-controllers and embedded memories. Usability in this context is defined as a quality attribute that assesses how easily a mitigation strategy can be implemented. Usability addresses questions related to integration, including the following: What do developers need to do to successfully configure and deploy a chosen mechanism? What is the limitation of the strategy, and is it possible to define the part(s) of the system that needs to be protected? Hence, the usability parameter of the strategy depends on three factors:

- The base system/device's properties.
- The requirements/limitations of the strategy.
- The non-functional requirements of the user.

For example, DECTED (Double Error Correction, Triple Error Detection) EDAC codes [7] show very good performance against soft errors, but integrating such approaches (hardware or software) into devices requires additional computational power as well as additional computing time, which are usually both limited in COTS devices.

V. Evaluation of Strategies

As shown in Section IV, it is crucial to estimate the performance and overheads of soft error mitigation strategies in order to identify appropriate strategies for one's problem domain given the underlying system requirements. This section demonstrates how to evaluate potential techniques in the context of an existing 1oo2D safety architecture. The 1oo2D safety architecture is considered fixed, and the goal of the evaluation is to provide the means to enhance this existing system regarding soft error mitigation.

The least complex solution (demanding only effort and time and no additional hardware) is to apply a software-based mitigation technique. However, the problem with software-based approaches is their memory consumption and computation time requirements as well as complexity of implementation.

Given an existing architecture which already provides certain features (e.g., 1oo2D safety architectures inherently provides redundancy, that can detect but not correct errors), a hybrid approach can be taken. In such an approach a system's existing features (e.g., error detection) are complemented with additional software-based mitigation techniques to achieve increased fault-tolerance (e.g., providing a correction mechanism to complement 1oo2D detection mechanism).

In the following, an analysis of the mitigation strategies explained in Section II will be presented, along with new ideas that utilize existing peripherals of the micro-controller. Techniques will be explained top-down to outline system safety-enhancement prospects for designers involved in different development stages of the system. In addition, the top-down order reflects the amount of effort required to implement a strategy in the system, as alterations in earlier phases might require a system redesign.

A. Software-based techniques

These techniques belong to the deployment phase according to Section II. While they do not require additional hardware per se, their overhead can affect the system's performance. To choose an appropriate strategy requires the comparative assessment of potential strategies. This section demonstrates how such an assessment could be performed via an exemplary calculation and comparison of memory consumption and runtime performances using the example of Parity Bit (PB) and Extended Hamming Code (EHC). A similar approach can be used when assessing other software-based mitigation strategies.

The evaluation is performed for varying lengths of protected data, as strategies scale differently with different lengths. For the representation of the codes, a common annotation (n, k) is used, where n denotes the number of total bits and k the number of protected data bits. The number of required check bits can be easily calculated as $n - k$. Utilizing these parameters, memory consumption (mc) is calculated in (1) and exhibited in Figure 14.

$$mc[\%] = (n - k)/k \cdot 100\% \qquad (1)$$

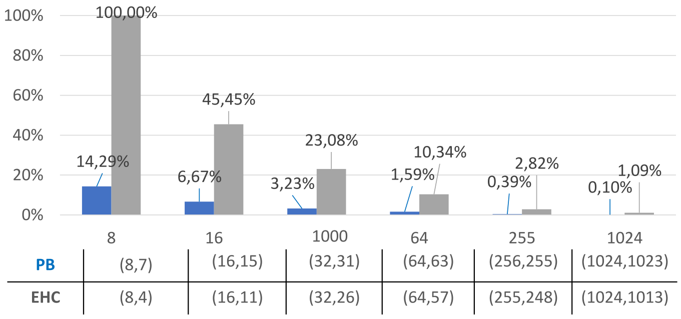


Figure 14. Memory overhead for different types of Parity Bit (PB) and Extended Hamming Code (EHC), where the x-axis denotes the total length of the word and y denotes the percentage of the memory overhead.

The run-time performance of a given strategy is closely connected to the complexity of the underlying algorithm. A good indicator of an algorithm's complexity is the number of logical XOR operators it requires for implementation.

In the context of PB, a calculation stemming from [50] was used. The algorithm is based on the consecutive application of shift and XOR operators. Alternatively, a lookup table could be used to calculate the parity bits of 8-bit words. While using a look-up table will slightly increase the memory consumption of the algorithm, it will decrease its complexity by 3 XORs.

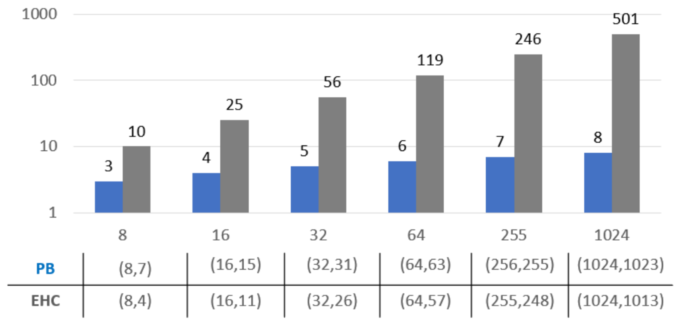


Figure 15. Number of XORs for encoding process for different types of PB(n, k) and EHC(n, k), where y-axis denotes the number of total XORs gates and the x-axis the number of the protected data bits.

Equation (2) was used to calculate the number of the XORs gates for EHC.

$$XORs(k) = 2^{k+1} - k - 3 \qquad (2)$$

where parameter k can be derived from the following form of Hamming code annotation $H(2^k, 2^k - k - 1)$. Equation (2) stems from [50], where it was calculated for the EHC recursive encoding computation. Figure 15 shows the number of XOR operators for varying lengths of protected bits.

PB and EHC differ significantly in terms of mitigation quality. While PB is only capable of detecting odd numbers of bit-flips errors (including single-bit errors), EHC can detect double-bit flips errors and correct only single-bit errors. In the context of safety-critical systems, this low mitigation quality will have a big impact on availability and safety.

In [51], a detailed report is presented on the number of soft errors in SRAM memory (512K x 8-bit) that were observed in space. Errors were recorded in a nanosatellite circulating the Earth's orbit. During the 2510 days of recording a total of 247593 soft errors occurred, which could be categorized into four types. The majority of the errors (i.e., a total of 244150 errors constituting 98.6% of the recorded errors) fell in the single-bit error class. Only 2996 errors (i.e., 1.21% of the recorded errors) constituted double-bit errors. Multiple bit (> 2) errors occurred at an even lower rate (corresponding to a total of 217 errors (0.08%)), while the remaining errors (230 (0.09%)) were classified as severe errors.

Let us consider the capability of the algorithms under test (PB and EHC) for this recorded error distribution. PB would detect all single-bit errors and some of the multiple bit errors, leading to a detection rate of 98.75%. PB detection alone is not enough and would not increase the availability of the system, because without recovery the sole identification of an error would lead to the system being put into a safe-state. Using EHC, 99.8% of errors would be detected and 98.6% would be corrected. This means that the system's availability could be increased significantly as it would only be stopped (put in a safe-state) for 1.4% of the errors. This leads to the conclusion that (on its own) EHC is significantly better when it comes to safety and availability, however, this can be associated with the higher memory overhead and complexity (as shown before). Furthermore, one should keep in mind that the SRAM used was relatively old (approximately 20 years), and thus, exhibits a lower probability for multiple bit errors because of the higher technology node in use. With newer memories utilizing smaller technologies, the distribution of the error is very likely to be different (i.e., more multiple-bit errors are to be expected).

In the context of safety-critical systems, the application of specific fail-safe architectures with hardware redundancy is very common. The next section will introduce a widely used fail-safe architecture and show how the application of simple EDAC codes can further improve a system's availability.

B. Hybrid techniques

While the previous section investigated purely software-based strategies, another option to increase a system's fault tolerance is to actively integrate the underlying architecture and components together with a software strategy.

In $1oo2D$ architectures with redundant memories (Section II-B), if an error appears it is not clear which memory was affected. Therefore, an error can be detected but not corrected and it will result in the system transitioning into a safe-state. A solution for overcoming this problem is to add mechanisms on top of the existing architecture that allow the recovery of faulty data and to extend the up-time of the system. Recovery mechanisms in this context are usually EDAC code based.

Adding additional hardware to the system is not feasible, as this would require redesigning the system, and an alternative option is to apply software-based EDAC code approaches.

1) Software Redundant parity: Given that 1oo2D already provides the possibility to detect memory errors, the question arises how existing architectures (i.e., 1oo2D) can be combined with software-based approaches.

A method for enhancing existing 1oo2D hardware architectures was proposed in our work [52]. This method constitutes an extension for mixed-critical real-time systems with an underlying 1oo2D architecture. We refer to it as Redundant Parity (RP). Figure 16 explains the basic concepts of the RP method. The method relies on 1oo2D's ability to detect soft errors and uses parity bits to establish the location of the error. Initially, the method generates parity bits for data that need to be protected (i.e., data in redundant memories). When bit flips occur and the 1oo2 comparator detects different bits in redundant data, the usual consequence is to generate a signal that will trigger the safe-state of the device. In contrast, our proposed RP method calculates new parity bits for both protected parts of the memories. In the next step, old parity bits are compared to newly calculated parity bits to establish the fault source. If the algorithm distinguishes between healthy and faulty data, the recovery phase is activated. Recovery is performed by simply overwriting the faulty data with the healthy data. Summarizing, the method uses the inherent capability of the 1oo2D architecture to detect bit flips. With the additional parity bit, the faulty redundant words can be determined and by means of redundancy, recovery is possible.

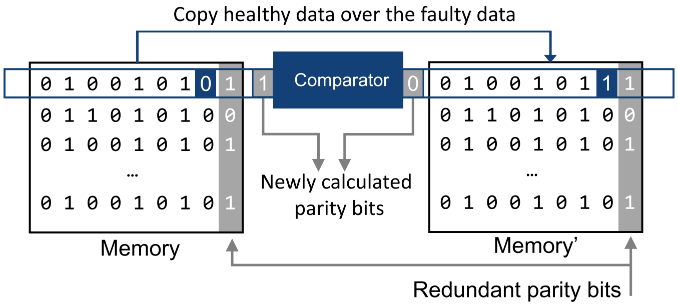


Figure 16. Redundant Parity method.

The method enables the correction of single-bit soft errors, which constitute the majority of soft-errors that occur. Odd multiple bit soft errors can also be corrected and even multiple bits can be detected. In the context of the recorded error data presented in Section V, this method would detect 100% of the errors and correct 99.4% of them. Memory overhead would be doubled and complexity would increase by twice the complexity of the parity bit.

Furthermore, the RP method provides separate detection and recovery phases, leading to less recovery time than in other EDAC methods. In addition, the proposed method is completely independent of the software architecture as it focuses on the memory's word level rather than on variables or structures [44]. However, the results also show that the application of the approach is limited to a 1oo2D architecture, which already provides the required data redundancy as well as self-tests to detect errors in the data.

2) Hardware redundant parity: As mentioned in section II, several affordable MCUs already integrate parity checks for SRAM memory. If HW-based parity checks are available, the Redundant Parity (RP) method explained in Subsection V-B1 could be implemented even more easily. This would help to overcome the main drawback of the RP method, i.e., an on-demand software-based calculation of the parity bit whenever a protected word is accessed in memory. Given the appropriate hardware, the calculation could be done automatically, minimizing the impact on the existing code. Using dedicated hardware would also relieve the CPU of the calculations required by PB. In case of discrepancy detection between redundant memories, the parity bit can easily be accessed and compared with the newly calculated parity bit, allowing a fast recovery procedure (i.e., overwriting a healthy over the faulty word) to be performed. While several MCUs (e.g., the STM32L4x MCU family) already provide inherent parity calculations, they often do not allow direct access to the calculated parity bits, which are calculated and saved internally. The only information provided by the system is a highly prioritized interrupt to the CPU, without any information about which of the memory addresses the error occurred in. A potential solution would be to scan the entire memory, but this not acceptable due to timing reasons and it would also defeat the purpose of using hardware.

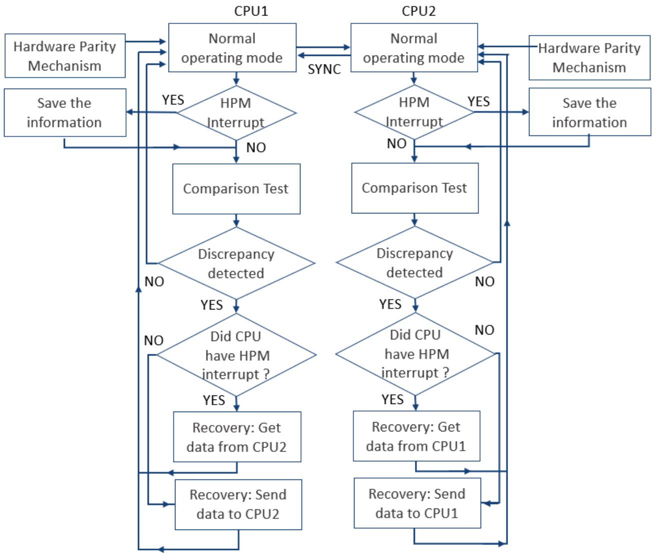


Figure 17. Flow chart diagram of Hardware Redundant Parity.

These limitations can be overcome with the following approaches: consider that a Hardware Parity Bit Mechanism (HPBM) detects an error in 1oo2 redundant memories. The CPU with the memory error will receive an interrupt. The information is saved, and CPUs continue with normal operations. Later, the 1oo2D comparison test detects a discrepancy between the redundant memories and its exact location. With the previously stored information about the location of the faulty memory (i.e., the saved interrupt), the fault can be pinpointed to one memory. If the interrupt is received on the CPU1 side, then data from CPU2 can be copied over the data of CPU1, otherwise, if CPU2 got an interrupt then data from CPU1 will be transferred to CPU2. If one of the CPUs gets more than 1 interrupt in the interval between two comparison tests, the safe-state should be activated, because the latest information about the faulty side will be wrong. Also, the safe-state will be activated if both CPUs receive an interrupt.

The flow chart diagnosis of this Hardware Redundant Parity algorithm is shown in Figure 17.

In contrast to its software-based predecessor, the approach does not require additional memory. Additionally, its complexity decreases as the overall code does not need additional changes, in contrast to the software RP algorithm, where the code to calculate parity bit needed to be inserted at every write request. The complexity of this approach is therefore very simple since no costly computations and no additional time is required, so overall system run-time is unaffected. In the context of safety, the functional safety assessment of the resulting system is easier. In other words, validation of the concept and showing that it has no false positives or false negatives is easier than in previous cases.

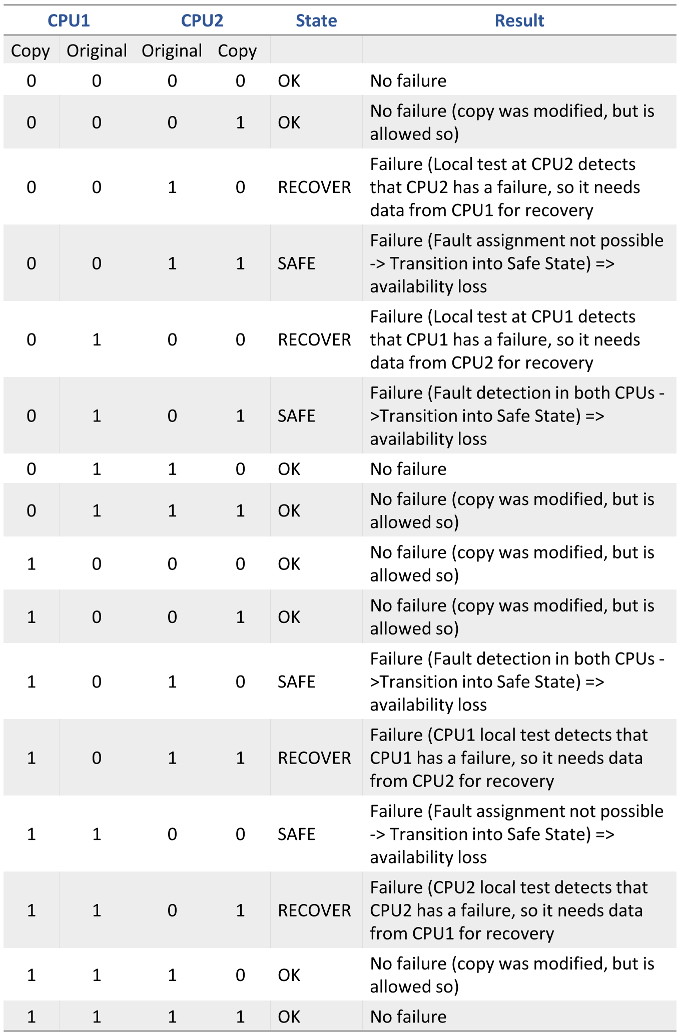
3) DMA based recovery: Direct memory access (DMA) is a method that allows an input/output (I/O) device to send or receive data directly to or from the main memory, bypassing the CPU to speed up memory operations. The process is managed by a chip known as a DMA controller (DMAC). The following mitigation strategy utilizes DMA method capabilities to protect memory with minimum changes to the operating code of the system.

Assume that a comparison self-test is done in slices as explained before. In the beginning, a copy of the safety-critical data is stored in the spare memory via DMA. As a result, both CPUs will have an original and a copy of the original safety-critical data. When an error occurs, i.e., a bit flip on the CPU1's memory, a comparison test will detect a discrepancy between the original data of two memories. Usually, this would lead to a safe-state but in this approach, recovery is possible and the safe-state can be avoided. After a discrepancy between memories is detected, each CPU starts a local self-test, comparing the original with the copied data. If the locally compared data is equal for CPU1 then data on CPU1 is intact and we can assume that the faulty memory is on CPU2. The recovery can be achieved by simply overwriting faulty data (CPU2) with healthy data (CPU1). If the locally compared data is not equal, then the assumption is that further corruptions occurred, and therefore, a safe-state will be activated. In general, the DMA method will theoretically cover all 1-bit errors. As shown in the example memory usage profile (Figure 12), although it is not possible to cover everything, a significant part of the memory will be covered. The previously described method's behavior for recovery and safe state handling, is depicted in Figure 18. The method can be applied in the same manner to CPU2.

The drawback of this approach is that additional memory is needed to hold copies of the data. This approach has a minor effect on the code because it only requires the configuration of the DMA and implementation of the recovery routine. The effect on the overall run-time is minimal because copying a few slices of the data should not have a significant impact. This method is not restricted to specific parts of the memories as in the case of HW redundant parity. Additionally, DMA is now a standard method that is included in most MCUs, therefore, it is not dependent on the MCU type.

C. Built-in hardware techniques

If none of the previous two categories fulfill the requirements for memory protection, then a redesign of the system should be considered. In this case, micro-controllers with built-in protection techniques should be used from the early design stages. These techniques are explained in Sections II-C and II-B. However, these techniques also have associated quality attributes, and thus, limitations have to be considered. For example, parity bit protected memories have a low mitigation quality (detection only), while ECC-Hamming code protected memories are better in this respect. But in some cases, run-time is affected or only some parts of the memory are protected. In general, when using built-in hardware, techniques will guarantee a better mitigation process but on the other hand, we are getting away from COTS MCUs and heading to safety-certified MCUs that are far more expensive. Nevertheless, as we stated in Section II-E, there are already some COTS MCUs availabe with built-in protection mechanisms. With advances in production techniques, we expect that the number of COTS MCUs with integrated measures will increase.



CPU1		CPU2		State	Result
Copy	Original	Original	Copy		
0	0	0	0	OK	No failure
0	0	0	1	OK	No failure (copy was modified, but is allowed so)
0	0	1	0	RECOVER	Failure (Local test at CPU2 detects that CPU2 has a failure, so it needs data from CPU1 for recovery
0	0	1	1	SAFE	Failure (Fault assignment not possible -> Transition into Safe State) => availability loss
0	1	0	0	RECOVER	Failure (Local test at CPU1 detects that CPU1 has a failure, so it needs data from CPU2 for recovery
0	1	0	1	SAFE	Failure (Fault detection in both CPUs ->Transition into Safe State) => availability loss
0	1	1	0	OK	No failure
0	1	1	1	OK	No failure (copy was modified, but is allowed so)
1	0	0	0	OK	No failure (copy was modified, but is allowed so)
1	0	0	1	OK	No failure (copy was modified, but is allowed so)
1	0	1	0	SAFE	Failure (Fault detection in both CPUs ->Transition into Safe State) => availability loss
1	0	1	1	RECOVER	Failure (CPU1 local test detects that CPU1 has a failure, so it needs data from CPU2 for recovery
1	1	0	0	SAFE	Failure (Fault assignment not possible -> Transition into Safe State) => availability loss
1	1	0	1	RECOVER	Failure (CPU2 local test detects that CPU2 has a failure, so it needs data from CPU1 for recovery
1	1	1	0	OK	No failure (copy was modified, but is allowed so)
1	1	1	1	OK	No failure

Figure 18. States of "DMA based recovery" operation within a 1oo2 memory architecture, regarding different perceived errors in CPU1 and CPU2's original data and copy data segments

VI. Conclusion

With decreasing transistor sizes, soft errors induced by external environmental factors increasingly constitute a problem for memory operation and provide challenges to ensuring a system's safety and availability.

The main goal of this work was to review mitigation strategies for 1oo2D safety architecture, which are applicable in different development phases of a system, as well as to identify the challenges which need to be considered during the design of soft-error mitigation strategies.

Today, several safety certified MCUs with integrated measures against radiation can be found on the market. COTS MCUs, on the contrary, are not always equipped with such protection measures and often only utilize the most simple protection techniques. As safety certified micro-controllers are becoming more expensive, industry often utilizes COTS micro-controllers in different safety architectures. These architectures rely on redundancy, i.e., the multiplication of systems, which can lead to even more expensive production costs and an increase in the overall system's complexity. Therefore, there is a need for solutions that utilize simple safety architectures together with additional techniques built on top of existing available architectures. Such approaches intend to keep safety at the demanded level but at the same time increase availability and reliability with minimal additional costs.

To increase availability and reliability within COTS memories, a certain level of fault tolerance is required. Current safety-critical applications rely on simple fail-safe architectures such as 1oo2D. The reliability and availability of fault-tolerant systems can be further improved, if such architectures are extended with additional software-based recovery techniques such as EDAC codes, which does not require additional hardware or a redesign of the underlying architecture.

As demonstrated in Section V potential mitigation strategies can be evaluated in terms of their overhead and complexity, as well as the different system development phases they apply to. Such a categorization of strategies highlights their individual cost and requirement trade-offs, their limits, and allows for the identification of suitable methods for specific application scenarios (e.g., when retrofitting existing brownfield automation devices).

When deciding on a method to be implemented on existing hardware, one must be aware of the associated overhead costs, as it will likely increase run-time and/or reduce available memory space. This aspect can be incorporated in strategy design by directly addressing mixed-criticality of data within correction and detection strategies, and differentiating among memory regions. This article demonstrated how such an assessment could be performed, by calculating and comparing memory consumption and run-time performances of different strategies, which can then be linked to the existing requirements of existing safety architectures, such as 1oo2D.

Software-based measures are rather difficult to use as they require implementation and integration into an existing system. If there is no other option, however, software-based measures must be implemented. In this case, two points should be considered: i) The usage of redundancy or coding theory (EDAC codes), where parameters such as quality and overhead (see Section IV) need to be taken into account. ii) The implementation has to be targeted at towards the usage profiles of the memory. Taking these profiles into account helps to reduce memory overhead and reduce implementation and integration overhead.

A thorough analysis of chosen strategies that are to be deployed in industrial controllers must be planned, in order to i) identify their limitations in the context of the system and ii) analyze the overall effect of the methods on the system regarding the associated challenges (see Section IV). Moreover, a detailed evaluation of a strategy's impact on a system's availability and reliability must be investigated in detail.

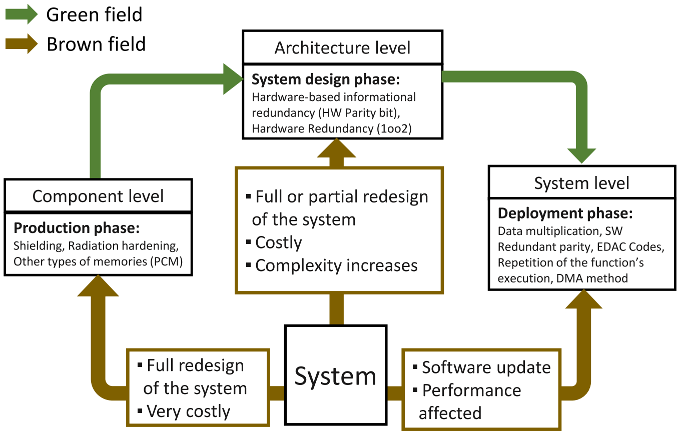

Green field
Brown field
Architecture level
System design phase:
Hardware-based informational redundancy (HW Parity bit), Hardware Redundancy (1oo2)
Component level
Production phase:
Shielding, Radiation hardening, Other types of memories (PCM)
▪ Full or partial redesign of the system
▪ Costly
▪ Complexity increases
System level
Deployment phase:
Data multiplication, SW Redundant parity, EDAC Codes, Repetition of the function's execution, DMA method
▪ Full redesign of the system
▪ Very costly
System
▪ Software update
▪ Performance affected

Figure 19. Deployment of mitigation strategies for greenfield and brownfield devices

A summary of this study's findings is presented in Figure 19. The green line presents different ways to mitigate soft-error for the different stages of system development. This option concerns greenfield systems (i.e., when designing a system from scratch). The brown line, on the other hand, represents options relevant to implementing additional soft-error mitigation strategies for brownfield devices (i.e., existing systems). Three approaches are possible for retrofitting brownfield automation. One is a complete redesign of the system, including measures such as shielding, hardening, or the selection and application of different, more resilient types of memory. This option might require more time and costs than expendable for an existing system. The second approach concerns a partial redesign, by adding additional components that increase the redundancy of the system. Although this approach is less expensive than a complete redesign, it is still associated with significant costs and effort. The last approach is to deploy software-driven approaches. While this approach is associated with the least costs, it requires extensive testing of non-functional parameters in order to make sure that the strategies are indeed applicable in the system context.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Austrian Research Promotion Agency (FFG) (#6112792).

REFERENCES

[1] A. Kajmakovic, K. Diwold, N. Kajtazovic, and R. Zupanc, "Challenges in mitigating soft errors in safety-critical systems with cots microprocessors," in PESARO 2020, The Tenth International Conference on Performance, Safety and Robustness in Complex Systems and Applications. IARIA, Feb. 2020, pp. 13–18.

[2] J. Vankeirsbilck, H. Hallez, and J. Boydens, "Soft error protection in safety critical embedded applications: An overview," in 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). IEEE, 2015, pp. 605–610.

[3] H. Iwashita, "International standards adopted by ITU-T to address soft errors affecting telecommunication equipment," ITU-T International Telecommunication Union - Telecommunication Standardization Sector, Geneva, CH, Standard, 2018.

[4] H. Forsberg and K. Karlsson, "COTS CPU selection guidelines for safety-critical applications," in 2006 IEEE/AIAA 25TH Digital Avionics Systems Conference, Oct. 2006.

[5] V. THATI, J. Vankeirsbilck, J. Boydens, and D. Pissoort, "Data error detection and recovery in embedded systems: a literature review," Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 3, 2017, pp. 623–633.

[6] M. Duncan and P. Roche, "Paving the way towards autonomous driving—tackling soft errors to security challenges," in 2017 IEEE International Reliability Physics Symposium (IRPS), 2017, pp. 2E–1.

[7] D. Elena, Fault-Tolerant Design. KTH Royal Institute of Technology, Krista, Sweden: Springer, 2013.

[8] A. Mukati, "A survey of memory error correcting techniques for improved reliability," Journal of network and computer applications, vol. 34, no. 2, 2011, pp. 517–522.

[9] E. Fujiwara, Code Design for Dependable Systems: Theory and Practical Application. New York, NY, USA: Wiley-Interscience, 2006.

[10] S. Jeon, E. Hwang, B. V. Kumar, and M. K. Cheng, "LDPC codes for memory systems with scrubbing," in 2010 IEEE Global Telecommunications Conference GLOBECOM 2010. IEEE, 2010, pp. 1–6.

[11] B. Tahir, S. Schwarz, and M. Rupp, "BER comparison between convolutional, turbo, LDPC, and polar codes," in 2017 24th International Conference on Telecommunications (ICT). IEEE, 2017, pp. 1–7.

[12] G. Mayuga, Y. Yamato, T. Yoneda, M. Inoue, and Y. Sato, "An ECC-based memory architecture with online self-repair capabilities for reliability enhancement," in 2015 20th IEEE European Test Symposium (ETS). IEEE, 2015, pp. 1–6.

[13] R. Santos, S. Venkataraman, A. Das, and A. Kumar, "Criticality-aware scrubbing mechanism for sram-based FPGAs," in 2014 24th International Conference on Field Programmable Logic and Applications (FPL). IEEE, 2014, pp. 1–8.

[14] M. Restifo, P. Bernardi, S. De Luca, and A. Sansonetti, "On-line software-based self-test for ECC of embedded RAM memories," in 2017 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT). IEEE, 2017, pp. 1–6.

[15] E. Peter and P. Salvatore, "Error detection in sram," Texas instruments, Application Report, 2017.

[16] N. Maruyama, A. Nukada, S. Matsuoka et al., "Software-based ECC for GPUs," in 2009 Symposium on Application Accelerators in High Performance Computing (SAAHPC'09), vol. 107, 2009.

[17] D. Dopson, "SoftECC: a system for software memory integrity checking," Ph.D. dissertation, Institute of Technology. Dept. of Electrical Engineering and Computer Science, Massachusetts, 2007.

[18] Intel® Embedded Memory User Guide, STMicroelectronics.

[19] MWCT101xS Safety Manual, NXP Semiconductors.

[20] F. Handermann, "Process safety architecture system neutral solution comparison," Chemical Engineering Transactions, vol. 48, 2016, pp. 499–504.

[21] R. Mariani and P. Fuhrmann, "Comparing fail-safe microcontroller architectures in light of IEC 61508," in 22nd IEEE International Symposium on Defect and Fault-Tolerance in VLSI Systems (DFT 2007). IEEE, 2007, pp. 123–131.

[22] S. Poledna, Fault-tolerant real-time systems: The problem of replica determinism. Springer Science & Business Media, 2007, vol. 345.

[23] OpenStax, Chemistry. Rice University: OpenStax, OpenStax Chemistry, 2014.

[24] F.-X. Yu, J.-R. Liu, Z.-L. Huang, H. Luo, and Z.-M. Lu, "Overview of radiation hardening techniques for ic design," Information Technology Journal, vol. 9, pp. 1068-1080, 2010.

[25] H.-K. Lim and J. G. Fossum, "Threshold voltage of thin-film silicon-on-insulator (SOI) MOSFET's," IEEE Transactions on electron devices, vol. 30, no. 10, 1983, pp. 1244–1251.

[26] T. Nakamura, H. Matsuhashi, and Y. Nagatomo, "Silicon on sapphire (SOS) device technology," Oki technical review, vol. 71, no. 4, 2004.

[27] M. Karagounis, D. Arutinov, M. Barbero, R. Beccherle, G. Darbo, R. Ely, D. Fougeron, M. Garcia-Sciveres et al., "Development of the ATLAS FE-I4 pixel readout IC for b-layer upgrade and super-LHC," Proceedings of the Topical Workshop on Electronics for Particle Physics, TWEPP 2008, Jan. 2008.

[28] F. Bedeschi, R. Fackenthal, C. Resta, E. M. Donze, M. Jagasivamani, E. Buda, F. Pellizzer et al., "A multi-level-cell bipolar-selected phase-change memory," in 2008 IEEE International Solid-State Circuits Conference-Digest of Technical Papers. IEEE, 2008, pp. 428–625.

[29] N. An, R. Wang, Y. Gao, H. Yang, and D. Qian, "Balancing the lifetime and storage overhead on error correction for phase change memory," PloS one, vol. 10, no. 7, 2015, p. e0131964.

[30] R. Forchhammer, "Automotive MCUs in28nm FD-SOI with ePCM NVM," STMicroelectronics, 2018.

[31] A.V.Kolobov and J. Tominagaand P.Fons, "Phase-change memory materials," in Springer Handbook of Electronic and Photonic Materials. Springer, 2017.

[32] L. Forbes, "Fully depleted silicon-on-insulator cmos logic," Dec. 14 2004, uS Patent 6,830,963.

[33] V. Technologies, "HARDSIL® integration & component design for foundries," MoPac Expressway, Suite 350, Austin, Texas, 7874, 2019.

[34] Product manual VA108x0, Vorago Technologies.

[35] "Iso 26262 - road vehicles – functional safety, part 1–10. electrical and electronic components and general system aspects," International Organization for Standardization, Geneva, CH, Standard, 2011.

[36] IEC, "International Standard 61508 Functional safety: Safety related Systems," International Electrotechnical Commission, Geneva, CH, Standard, 2005.

[37] F. Nocha, "Ecc handling in tmsx70-based microcontrollers," Texas instruments, Application Report, 2011.

[38] F. Reichenbach and A. Wold, "Multi-core technology–next evolution step in safety critical systems for industrial applications?" in 2010 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools. IEEE, 2010, pp. 339–346.

[39] C. Preschern, N. Kajtazovic, and C. Kreiner, "Built-in security enhancements for the 1oo2 safety architecture," in 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). IEEE, 2012, pp. 103–108.

[40] STM32F4 Series safety manual - user manual, STMicroelectronics.

[41] C.-W. Wu, "Chapter 8 - memory testing and built-in self-test," in VLSI Test Principles and Architectures, L.-T. Wang, C.-W. Wu, and X. Wen, Eds. San Francisco: Morgan Kaufmann, 2006.

[42] A. J. Van De Goor, "Using march tests to test srams," IEEE Design Test of Computers, March 1993.

[43] Handling of soft errors in STM32 applications, Intel.

[44] A. Kajmakovic, R. Zupanc, S. Mayer, N. Kajtazovic, M. Hoeffernig, and H. Vogl, "Predictive fail-safe improving the safety of industrial environments through model-based analytics on hidden data sources," in 2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES). IEEE, 2018, pp. 1–4.

[45] A. Burns and R. I. Davis, "Mixed criticality systems - a review," in Department of Computer Science, University of York, York, UK, 2015.

[46] J. S. Miguel and N. E. Jerger, "Data criticality in network-on-chip design," in Proceedings of the 9th International Symposium on Networks-on-Chip, 2015, pp. 1–8.

[47] L. Botler, N. Kajtazovic, K. Diwold, and K. Römer, "JiT fault detection: Increasing availability in 1oo2 systems just-in-time," in Proceedings of the 15th International Conference on Availability, Reliability and Security, ser. ARES '20. New York, NY, USA: Association for Computing Machinery, 2020.

[48] K. Itoh, "Embedded memories: Progress and a look into the future," IEEE Design & Test of Computers, vol. 28, no. 1, 2011, pp. 10–13.

[49] Reference manual for STM32 applications, Intel.

[50] L. Zhengrui, L. Sian-Jheng, and H. Honggang, "On the arithmetic complexities of Hamming Codes and Hadamard Codes," 2018.

[51] H. Caleb and B. Vipin, "Error detection and correction on-board nanosatellites using Hamming codes," Journal of Electrical and Computer Engineering, 2019.

[52] A. Kajmakovic, K. Diwold, N. Kajtazovic, R. Zupanc, and G. Macher, "Flexible soft error mitigation strategy for memories in mixed-critical systems," in 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). IEEE, 2019, pp. 440–445.

Investigation of Problems with High Initial and Update Efforts in the Modeling of Production Systems

A Review on System Modeling Approaches

Marius Heinrichsmeyer[1], Amirbabak Ansari[2], Nadine Schlueter[3], Christian Boehmer[4]
Product Safety and Quality Engineering
University of Wuppertal
Wuppertal, Germany
E-Mail: heinrichsmeyer@uni-wuppertal.de[1], aansari@uni-wuppertal.de[2], schluete@uni-wuppertal.de[3], christian.boehmer-hk@uni-wuppertal.de[4]

Abstract — **The need to use Model-Based Systems Engineering (MBSE) has seen an upswing, especially in recent years, for example, due to the ever-increasing complexity of products and production systems. Nevertheless, evaluations of the current state of research and experience from our own completed and ongoing DFG projects (KAUSAL, ReMaiN, and FusLa show that the use of MBSE in the industry is underestimated, mostly because of the enormous initial and update efforts in the modeling. Approaches that support system modeling, such as Modelica, SysML, and eDeCoDe or approaches for their partial automation only help to a limited extent to reduce the modeling effort when mapping production systems. For this reason, the research group of Product Safety and Quality (PSQ) intends to research possibilities and opportunities for partial automation in the modeling of production systems. To achieve this, the problem of excessive initial and update efforts when using MBSE explicitly in the modeling of production systems should first be highlighted and developed as research potential.**

Keywords-Model Based Systems Engineering; Partial Automation; Failure Cause Localization; Production.

I. Introduction

Following our paper in ICONS 2020 about the validation of a Failure-Cause Searching and Solution-Finding Algorithm (FusLa) in production, it was stated that a detailed production system model forms the basis of localization of failure causes [1]. In this paper, the initial effort of system modeling and its updating is investigated.

System models can be used for many purposes, including the visualization of production systems. They are particularly important in order to master the increasing complexity of product and production systems as part of MBSE [2][3][4]. As a simplified representation of a complex system, system models form the basis for the design and improvement of processes according to failures and previous analyzes. However, the initial and update effort for creating a system model and the effort for the introduction and application of systems engineering is enormous, since companies have to use many tools or toolchains to be able to correctly map the complex information [5]. This effort shows itself particularly in high personnel costs and a considerable amount of time expenditure. In the coming years, a further increase in the resources, which are required for modeling the production system, is to be expected. It is because of the increasing number of components and their connectivity with each other and also the increasing variety of requirements, while the development and testing times for products or production systems are reducing [6]. Existing approaches to partial automation of the creation of system models are very specific and only consider just some aspects of the overall system, such as the requirements [7]. So, they cannot be used for a holistic system description. To reduce the initial effort for the creation and then the maintenance of a system model for companies and to reduce the resource expenditure, it is necessary to develop a practicable and scientific approach, with which systems can be modeled partially automated based on existing documents and information. However, in order to be able to implement such a development, three key questions need to be asked:

1) How does the modeling of a production system work?

The second section of this paper looks at how a production system can be represented as a model, and which elements are necessary for this. This is necessary since there are various considerations regarding the representation of models. Some approaches consider production systems as the interaction of the subsystems, while others consider inputs and outputs as well. Section II is primarily intended to describe the different forms of modeling of production systems and to specify their use cases.

2) Which approaches contribute to the modeling of a production system and how much effort is required?

Based on the modeling forms, the next step is to question, which approaches to modeling are already available and how they contribute to the mapping of a production system. This will not only indicate the limits of existing approaches regarding the modeling of production systems but also show the initial and update effort associated with their modeling. Overall, this makes it possible to determine a statement, to what extent the mentioned problems and efforts are already compensated or intensified by existing approaches.

3) Which approaches already contribute to the reduction of the initial and update effort. Are these sufficient?

In the last step, based on the initial and update efforts of each approach, it is then examined, which existing approaches already contribute and can contribute to the reduction of mentioned efforts. This step will provide a statement about whether current approaches are sufficient to eliminate the mentioned problem, or whether there must be further research projects and new approaches to be developed, which can contribute to an elimination of the problem.

To investigate these questions, Section II gives an overview of the types of system modeling. Section III discusses the state of the art in modeling approaches that deal with standardized modeling of systems and Section IV discusses those that contribute to partially automated modeling. Finally, Section V gives an overview of the research topics to be pursued.

II. MODELING OF PRODUCTION SYSTEMS

In the literature, there are numerous definitions of the term model, which originate from different industries and fields of application. What they have in common is that a model is an abstract representation of reality [8]. The systematic creation and the integrated use of digital system models in the context of the MBSE serve the purpose of making the increasing complexity of products and processes manageable [3][9]. However, how is the modeling of a production system accomplished?

Remarkably similar to the concept of the model, the related process of modeling is also defined in many ways. For the modeling of production systems, however, the modeling focuses on three main forms of representation, including functional, hierarchical, and structural modeling. Which form of presentation is most suitable depends largely on the object under consideration and the application [10].

1) *Functional modeling*

The functional form of modeling considers a model at the top level. As shown in Figure 1, this form of modeling models a production system as an operational conversion and transformation process, by which a set of outputs (e.g., products or services) is created from a set of inputs (e.g., material, energy) through the work of human and/or the use of work equipment [11]. This form of modeling is particularly suitable if a holistic view of the production system concerning other systems, such as product development or top-level use, should be achieved over the product life cycle [10].

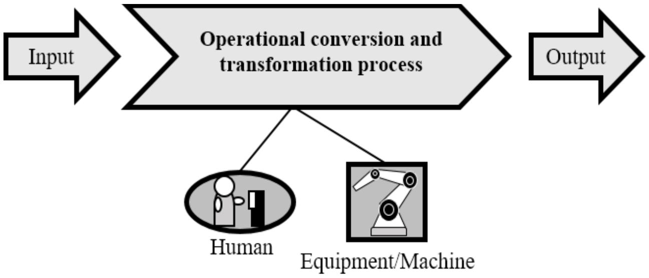


Figure 1: Functional modeling form of a production system.

2) *Hierarchical modeling*

The second form of representation of the modeling is called hierarchical modeling and covers production systems via subordinate and superordinate subsystems. In contrast to functional modeling, in which the highest level of detail is considered, hierarchical modeling already shows the first relationships between subsystems in more detail. This form of modeling is particularly suitable when the interaction of higher-level processes in the production system, e.g., purchasing or manufacturing, is to be analyzed. Above all, the recording of material and information flow is possible with this form of modeling [10].

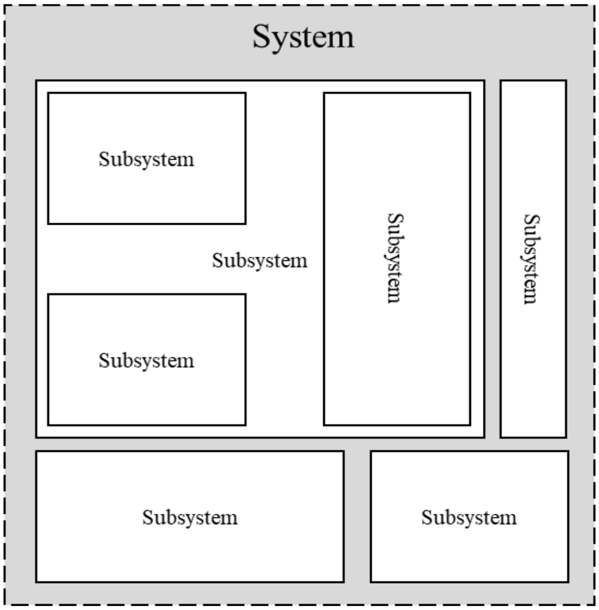


Figure 2: Hierarchical modeling form of a production system.

3) *Structural modeling*

Structural modeling represents the last form of modeling of production systems. Here, the production system is divided into different components, including system elements, their relations, inputs, outputs, the system environment, and the system boundary. This is the most detailed form of modeling. This is particularly suitable for understanding the interrelationships between different system elements and making the complexity of a holistic production system more manageable. In addition, this amount of detail makes it possible to ensure the traceability of system elements by evaluating their relationships [10].

As already mentioned, the selection of a suitable form of representation of the modeling largely depends on the object under consideration and the application. This suggests that the elements that are required to map a standardized production system model also vary on a case-by-case basis. However, experience from previous fundamental research projects, such as KAUSAL and in part, ReMaiN, showed that the structural modeling form, in particular, can be classified as suitable when it comes to analyzing and understanding the interrelationships within production.

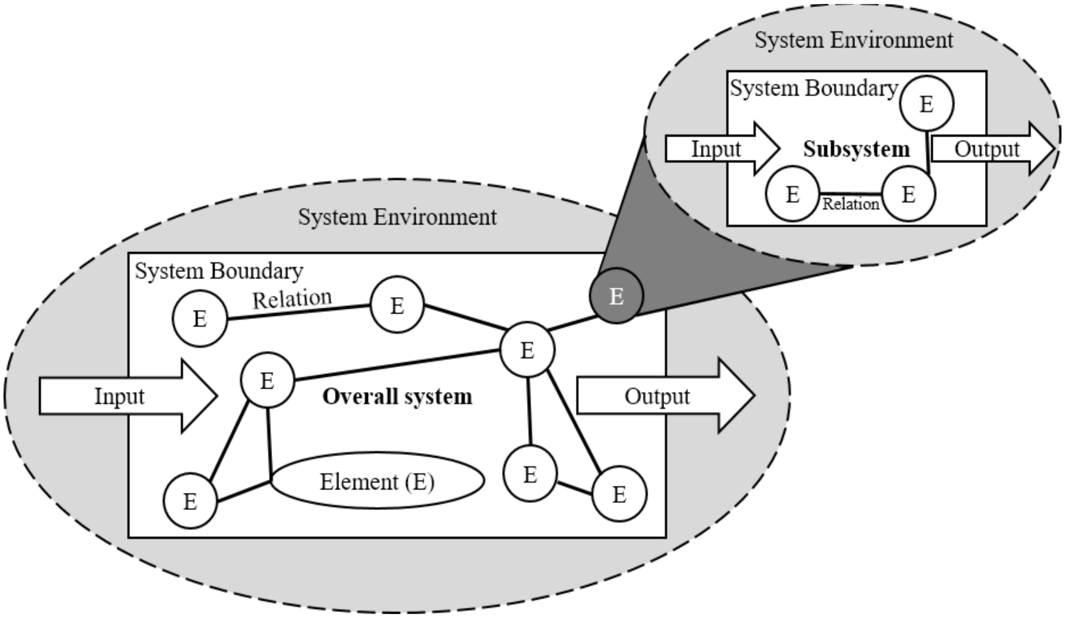


Figure 3: Structural modeling of a production system [12].

Nevertheless, this is also to be assessed disadvantageously, since the high level of detail of the structural modeling also entails an enormous challenge for the companies. The challenge is particularly noticeable in the initial and update effort already mentioned. Figure 3 demonstrates an exemplary structural system model and its elements

Specifically, structural modeling means that every system element, be it a machine, a person, or the input and output, must be recorded and related. Especially with extremely complex production systems, such as those found in the automotive industry, such modeling could hardly be carried out by individual people. Instead, individual partial models from different areas are developed. However, these are designed for a specific problem and do not help to understand the holistic production system model in detail. In order to counteract this problem and to simplify the modeling itself, different modeling approaches have been established in recent years. These specify which system elements are to be classified as necessary for the modeling and how their interrelationships are to be understood. The main aim of these approaches is to make the complexity of the production systems more manageable through suitable and, above all, less complex modeling.

To evaluate these approaches regarding their suitability concerning the modeling of production systems and their effort, some established approaches are presented below and critically examined. The subject of consideration is structural modeling, since, as already mentioned, this involves the greatest initial and updating effort.

III. Approaches to Mapping the Standardized Production System Model

Approaches that are considered in the context of the contribution are Modelica, CONSENS (Conceptual design Specification technique for the Engineering of Complex Systems), SysML (Systems Modeling Language), MES (Manufacturing Execution System), and Demand Compliant Design (DeCoDe). The initial and update effort was assessed after practical application of the respective approaches and is summarized using the assessment scheme ● = high effort, ◑ = moderate effort, and ○ = little to no effort.

A. Modelica

The first approach, “Modelica”, enables object-oriented modeling of complex heterogeneous systems. For this purpose, a description defined by a language code is translated using hierarchical object diagrams specified by a library. The interrelationships between the elements must always be physical [13][14].

Modeling with Modelica has both advantages and disadvantages. On the one hand, it is a simulation tool that enables the quantitative analysis of system behavior within the usage phase. A combination with other methods such as Fault Tree Analysis or Markow models can be implemented and the visualization also is not limited to a single medium. On the other hand, only the component view is considered in the visualization. Therefore, a statement regarding the involved functions, processes, and requirements cannot be made. This in turn means that the traceability of failures cannot be guaranteed. Regarding the effort involved in structural modeling, it was found that this, of course in direct comparison with other approaches, should be assessed with a moderate effort (◑). The background of this assessment lies in the focus on the component view. While other approaches consider other system elements, such as requirements or processes, and also take their interrelationships into account, the model with Modelica captures only one type of system element.

B. CONSENS

CONSENS is a specification technique used to describe the principle solution of mechatronic systems and the associated production system [15]. With this approach, ten partial models are defined, seven of which, as shown in Figure 4, describe the problem solution (environment, application scenarios, requirements, functions, active structure, shape, and behavior) and the remaining three (processes, resource, and shape) describe the production system. The language uses a visual syntax and since the semantics are already defined, it can be used effectively without any adjustments. This can also be extended via profiles [16].

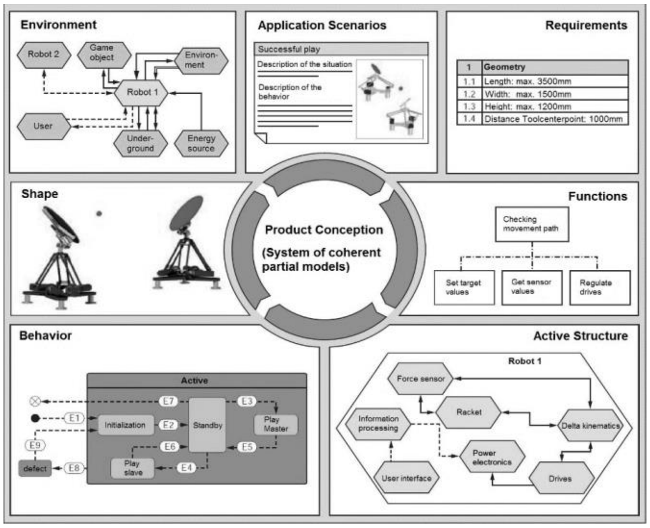


Figure 4: CONSENS approach according to [17].

In this model, requirements are listed, classified, and connected with the functions and system elements. The structure and the mode of action are represented by the structure of action, the core of the model [18].

One advantage of this model is that it forms a basis for discussion and documentation, especially in the planning phase. On the other hand, there is a connection between the views of the requirements, functions, and components. In comparison to Modelica, CONSENS records the behavior of the system model with the help of application scenarios. A disadvantage is that although there is a network being formed, there is no consideration of its interrelations. Besides, due to the numerous and, above all, extensive diagrams, the overall model quickly becomes confusing and even more complex. It should also be added that traceability is only partially guaranteed with this model. Regarding the initial and update effort with CONSENS, one can see that the modeling is of high effort (●). The acquisition of all system elements via the corresponding partial models as well as continuous updating by changes to the system are extremely resource-intensive. Above all, taking system behavior into account via corresponding application scenarios can be classified as a great effort, since the scenarios have to be individually adapted to the respective production system models.

C. *SysML*

SysML is a modeling language based on the Unified Modeling Language (UML) [19]. In contrast to CONSENS, SysML visualizes additional elements (e.g., requirements and functions) and offers a modeling of use cases as well as further possibilities [19]. It has its own notation so that system elements and relationships can be assigned. SysML is widely used because it is highly extensible and adaptable to the respective development task, e.g., through ready-made profiles [20]. However, adaptability is also necessary, since the semantics contained in SysML are only rudimentary compared to less frequently used alternatives [16].

As shown in Figure 5, the system model is characterized by various diagrams (e.g., diagrams of structure, behavior, requirements, parameters, and use cases) [21].

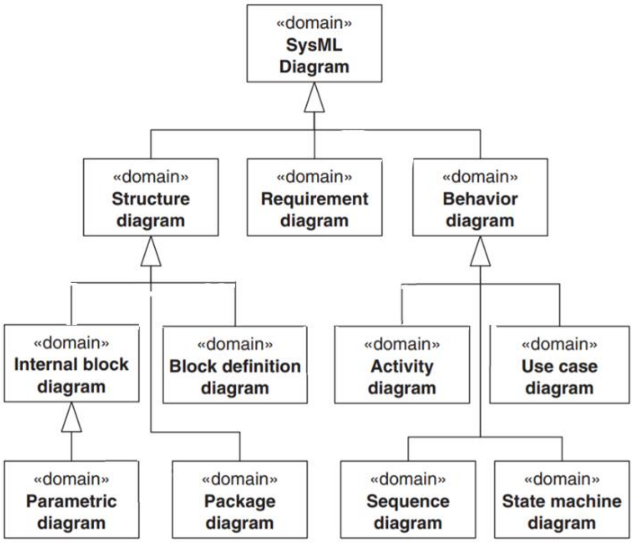

«domain» SysML Diagram
«domain» Structure diagram
«domain» Requirement diagram
«domain» Behavior diagram
«domain» Internal block diagram
«domain» Block definition diagram
«domain» Activity diagram
«domain» Use case diagram
«domain» Parametric diagram
«domain» Package diagram
«domain» Sequence diagram
«domain» State machine diagram

Figure 5: SysML diagrams according to [21].

The relatively large number of diagrams makes it possible to visualize the system model from different perspectives. At the same time, however, this is also a disadvantage, since the enormous number of diagrams and their defined structures do not allow intuitive use [19]. In addition, SysML was originally used in software development and later adapted for product development and is therefore not suitable for modeling production systems.

The application of SysML also involves a high effort (●). Although SysML can be simplified by supporting software systems such as Cameo Systems Modeler, numerous diagrams must be worked out and related to each other. The advantage of SysML, but not the decisive factor, is that the system elements are available across the diagrams. This means that when an explicit system element is changed, all system elements with the same identifier will also change. Above all, this reduces the update effort, since not every system element has to be changed individually.

D. *User-oriented System Modeling*

Florian Munker presents in [22] his approach to user-oriented system modeling. It aims at developing a concept that allows an easy entry into interdisciplinary system modeling while maintaining agility and flexibility. By determining boundary conditions and based on different approaches investigated, a user-oriented and integrated initial approach was developed, which should consist of language, method, and tool. This approach was then presented at the Systems Engineering Day 2015 [23].

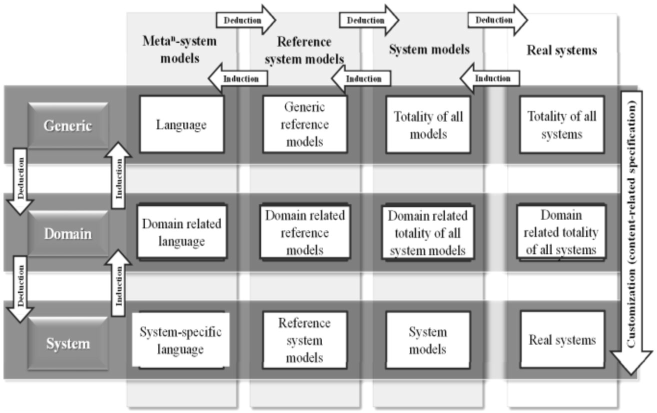


Figure 6: Abstraction of the system model [22].

As shown in Figure 6, a framework should be used, which should enable access to essential information on past product generations by accessing older system models. Thus, the modeling effort shall be reduced by transferring this information. The prototype worked was then translated into program codes by an assistant within one year. The different modules include the reading of the project and the Metadata, the graphical representation, the processing of the information according to the user stories, which served as requirements, and the saving of the data. However, the prototype was developed with some limitations, so that only the boundary conditions identified as mandatory were considered. These include user-oriented object modeling, graphical modeling, and view generation. The restrictions are thus "essentially the limitation of the realization to the partial model 'system structure' so that a fundamental system modeling can be tested on it [22 p. 70]". Furthermore, the modeling of the remaining partial models has been simplified. The created prototype was then used and evaluated by a test group. Part of the application study was also the Graphical User Interface, which is divided into a graphical modeling interface, buttons for modeling partial models and features, an administration area, and an area for the structure trees of the partial models. The bottom line is that this type of modeling also involves a high effort (●) and the suitability in practical application is rated as low, while the necessity of such an application and the potential of this prototype are confirmed.

E. MES

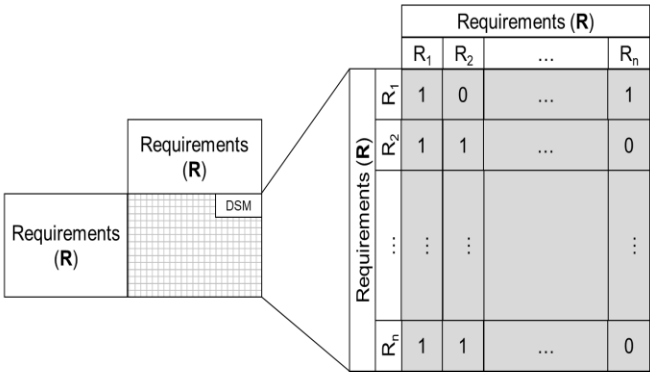
Another approach that is already established, especially in industry, is MES. These systems form an interface between the planning systems used, including ERP for example, and the equipment or personal interfaces present in the production systems. MES systems are primarily used to capture all processes in production systems, e.g., which equipment produces which product, to process them in real-time and to control them accordingly. This makes it possible to determine the process capability of running processes throughout the entire production system and to initiate measures to restore process capability in case of any deviations. In addition, material bottlenecks, e.g., in value creation with suppliers, can also be detected and compensated for at an early stage. The corresponding modeling of MES can vary depending on the company. While some companies embed CAD models of the facilities into the production system model, other companies only consider data evaluation [24]. Overall, however, it can be said that MES is quite capable of capturing corresponding processes, facilities or requirements to be implemented. However, an extensive acquisition of the persons including their competencies is missing. This has the background that MES systems are currently not yet developed for the optimization of people in the production system model, but focus primarily on the optimization of the process view [25].

The effort of MES modeling, especially concerning the initial implementation, is a challenge for companies. To be able to work with MES, companies must have a corresponding infrastructure within their production system. This means that the data of the facilities and machines must also be accessible and personal interfaces must be available. If this is not the case, massive intervention in the actual production system is required first. For this reason, the effort involved in dealing with MES is also classified as very high (●).

F. eDeCoDe

eDeCoDe is an approach for the standardized description of a sociotechnical system model under the principles of systematical thinking and acting [26][27]. The eDeCoDe model is used to mentally decompose sociotechnical systems into five different views. These include requirements (R), functions (F), processes (P), components (C), and persons (Pe) of the system under consideration. These views are arranged in the form of matrixes, which are linked to each other. There are also some tools and questions that are provided to help capture these links. eDeCoDe is a procedure for creating a transdisciplinary system model [26].

The eDeCoDe tools, including the Design Structure Matrix (DSM), Domain Mapping Matrix (DMM), and Multi-Domain Matrix/Multi-Domain Graph (MDM/ MDG), statically map the technical system under analysis. By adding the fifth view, the eDeCoDe tools also make the modeling and investigation of sociotechnical systems possible. The DSM allows the qualitative capture of different elements of the same view (e.g., functions).



	R_1	R_2	...	R_n
R_1	1	0	...	1
R_2	1	1	...	0
⋮	⋮	⋮		⋮
R_n	1	1	...	0

Figure 7: The DSM Matrix (Requirements View) [27].

As shown in Figure 7, by listing all elements equally on the axes of the square matrix, interrelationships between the elements can be identified using the notation 1 = relationship, 0 = no relationship [26] [28] [29].

The DMM is an extension of the DSM. While the DSM only considers the elements of a single same view, the DMM comprises the elements of two different views (e.g., functions and requirements). This makes it possible to capture the interrelationships between the elements of the views and thus also to link the views with a visualized notation [30].

As shown in Figure 8, the combination of DSM and DMM is called MDM. Similar to DSM, MDM is also a square matrix with equal axes, but this time it captures all views (requirements, functions, processes, components, and people), elements, and relations. By representing each element of the system through the views, it enables the derivation of indirect dependencies of the system elements under consideration.

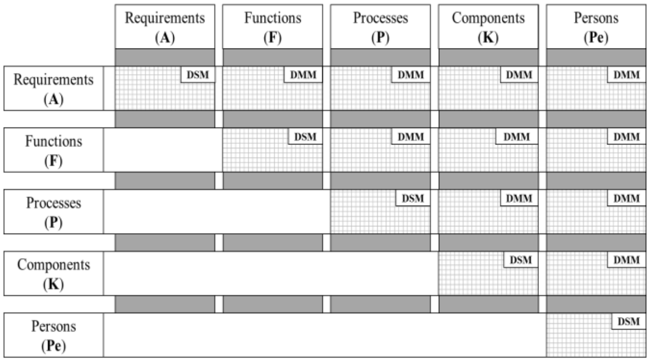


Figure 8: Combination of eDeCoDe Matrixes according to [27].

An advantage of the DeCoDe tools is that the system does not have to be completely mapped before it can be analyzed and designed [26]. This results in a system model that is reduced in complexity, although according to [31], this is associated with increased environmental complexity for this system. Furthermore, it is possible to illustrate the results resulting from the matrices in the form of graphs, so that the understanding of complex issues can be simplified by this kind of modeling [30].

The application of eDeCoDe has also proven to be extremely complex (●). The background of this is that each system element must first be worked out separately and then, in a further step, they will be related to each other. With extremely complex and continuously changing production systems, this task seems to be almost impossible to be accomplished by individual employees. Similar to SysML, eDeCoDe can also be supported by appropriate software in the actual process, such as LOOMEO. However, the initial and update effort remains almost identical.

After evaluating the corresponding efforts by applying the respective approaches to structural modeling, the result seems to show clearly that Modelica, in terms of effort, seems to be the most suitable. Nevertheless, at this point, it is necessary to critically question whether Modelica is sufficient to describe a production system model holistically since it only represents the component view. So capturing of interrelated processes or requirements is completely absent. However, this is necessary if an analysis of the facts within the production system is to be carried out. For this reason, the evaluation allows the statement that Modelica is not sufficient to model a production system, despite the lower initial and update effort required for modeling. However, which of the approaches then seems to be the most suitable concerning the respective effort involved?

G. *Which approach is best suited for modeling a production system model?*

As already mentioned, structural modeling varies according to the consideration and application of the model. Therefore, to answer the above question, it must first be clarified, which object of consideration and application is involved. These can always be different. Thus, the production system model can be used to evaluate the information flows regarding data protection or to identify the causes of failures in the model based on detected failures in the use phase. Despite the variation of the objects of consideration and use cases, the literature shows that a model of a production system can be considered from five standardized views [32]. The five views, visualized in Figure 9, are a superordinate grouping of the individual system elements of the model. These include requirements (R), processes (P), people (Pe), functions (F), and components (C). According to [26], these views are necessary to represent a sociotechnical system, including a production system, in its entirety [33]. Besides, these views enable the traceability of individual system elements via the interrelationships within the production system model.

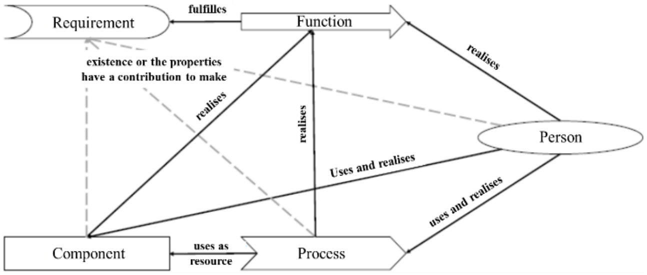


Figure 9: Interrelationships of system elements in eDeCoDe [27]

Based on this prerequisite, the eDeCoDe approach offers the greatest potential for structural modeling of a production system model. Despite the possibilities of eDeCoDe, it has already been shown that this approach involves an enormous initial and update effort. Therefore, the eDeCoDe approach is certainly suitable to make the complexity of a production system more manageable. Nevertheless, its modeling poses a great challenge to companies in terms of the effort involved. To compensate for this challenge, approaches were researched and evaluated, which can contribute to the partial automation of eDeCoDe modeling. The aim was to investigate whether partial automation of such a modeling is already possible, or whether the problem of excessive initial and update efforts in the modeling of production systems still exists.

In order to evaluate these approaches about their suitability for the partial automation of the modeling of production systems and concerning their limits and effort, some established approaches are presented and critically questioned in the following. The object of consideration is the structural modeling with eDeCoDe, since this, as already mentioned, involves the greatest initial and updating effort.

IV. Approaches to Partially Automated Modeling

In the literature, some individual approaches can be used to support the modeling of production systems. Especially the aspect of partial automation is considered to have great potential. Approaches that are included are, e.g., AAES and ARIS, which are described in detail below.

A. *AAES – Requirements View*

AAES is a method, with which the step from document-based to model-based requirements engineering (RE) as a starting point for MBSE is facilitated. With this method, specifications can be automatically broken down into individual requirements, which are subject to comprehensible versioning and are efficiently transferred to RE tools [34]. Finally, this also serves to quickly evaluate new requirements and initiate the implementation of these. Thus, the efficiency can be increased and at the same time, an increased acceptance of the changes by the users can be achieved. The starting point for the development was that many requirements are currently still stored in text-based documents that cannot be read by MBSE tools. If these continuous texts are now to be transferred to RE tools or modeling tools, this would mean that all requirements would have to be transferred manually. According to [34], this would go hand in hand with reduced quality and speed of the transmission, reduced profitability, and reduced user acceptance and motivation. However, since AAES can automatically read PDF-based documents, such as the specifications document, and forward them in ReqIF format to RE tools, which in turn can be linked to modeling tools, these effects can be counteracted preventively.

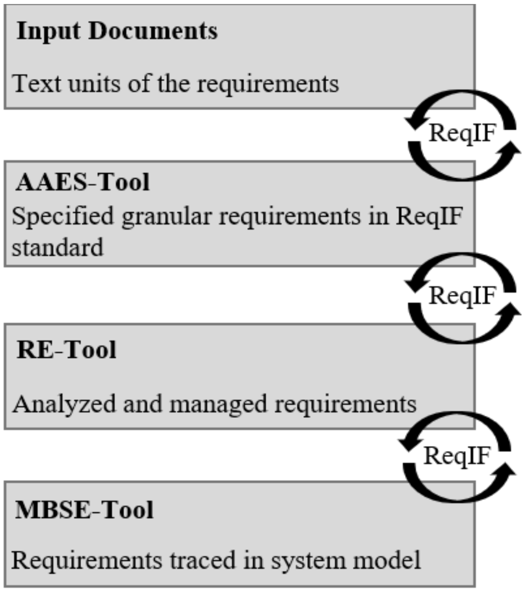

Input Documents
Text units of the requirements
ReqIF
AAES-Tool
Specified granular requirements in ReqIF standard
ReqIF
RE-Tool
Analyzed and managed requirements
ReqIF
MBSE-Tool
Requirements traced in system model

Figure 10: Process of requirements work from text to system model according to [34].

Besides, due to the growing complexity and its dimensions of variety, connectivity, dynamics, and globalization, a company must be able to act agilely and flexibly and at the same time guarantee traceability [35]. This means the networking of requirements with the product structure, tests, and the "atomic requirements gathering" are more relevant than ever [26][34]. The prerequisite for AAES is that requirements documents must be a structured set of data created and stored as a unit. If this requirement is met, the process of transfer based on the INCOSE manual or the phases of the V-Modell can be initiated. First, stakeholder requirements must be defined for this purpose, followed by a requirements analysis. This is followed by the architecture design, the design definition, and finally the system analysis.

B. *Analysis-simulation Models*

This approach is intended to contribute that reduces the manual effort required for simulation-based analyses. System simulations combined with fault injections can be used, for example, to support an FMEA, i.e., to assess the reliability of systems. Such a procedure is also recommended in ISO 26262:2015 "Functional safety of motor vehicles" to estimate the achieved Automotive Safety Integrity Level (ASIL). Model-Driven Development techniques are used for the specification of the failure effect simulation so that the effort of the failure effect simulation can be reduced by automated code generation and efficient reuse of simulation models using a component library. The effort of documenting the analyses according to ISO 26262 is also reduced. The UML profiles are also used because some extensions such as SysML and MARTE are already established in the automotive industry [36]. The connection to existing modeling languages is done with Model-to-Model transformation techniques (M2M). Furthermore, code generation techniques are used to automatically generate the structural part of the program code from the class descriptions: The code is highly reusable so that only the functional part of the code has to be added manually. A kind of top module instantiates, configures, and links the models of the simulation. The linking of the analysis results with the specifications of the system models can be done in two ways, semi or fully automatic.

This approach is also pursued in other methods. For example, there are overlaps with the method described in [37]. This approach presents a method of automatic generation of simulation models for production planning. It allows the automatic generation of simulation models of production systems based on data from the production planning and control system (PPC system). Thereby methods of data mapping, data transformation, data storage, and an intermediate data model are used. Thus, the effort for simulation projects, which accounts for about 30-40% of the total duration of data collection and up to 35% of model preparation exists, can be reduced [37]. The aim is to prevent serious failures in the design phase of the model by determining restrictions, definitions, and structures.

C. MDSOA

The approach of "model-driven service-oriented architecture" describes the use of different methods and notations to refine models through automated model transformations and the generation of artifacts [38]. MDSOA can be applied to any software development process. It uses model transformations to automate recurring tasks. Among other things, the quality assurance process can be automated.

The approach is based on the OMG's MDA standard for model-based software development and is similar to the "modeling and simulation as a service" (MSaaS) approach presented in [39]. It also introduces automated model transformations that should enable users to model in their languages. The model-to-text transformation is the core of the model-driven process: the generator model serves as input; the output is the memory library in JavaScript and an HTML file that ensures the actual implementation.

D. Machine modeling – component view

According to [40], the effort for creating the machine model in simulation projects is often higher than the benefits derived from it. To prevent this effect, a method was developed, with which a machine model can be created automatically from the engineering documents. No detailed knowledge of the machine is necessary, and consequently, no expert has to be involved in creating the machine model. The modular approach used in the Aquimo project automatically configures interdisciplinary engineering documents and the machine model. Among other things, company and project-specific parameter values and the installation diagram are used for this purpose [41]. The behavior models of the components are also created automatically. Another approach is the approach by Reinhart et al. presented also in [40], in which a meta-model is created and the interfaces of the required modules are manually coupled. The subsequent parameterization is also done manually, while the machine model is generated in a partially-automated manner. As shown in Figure 11, the approach in [40] itself uses the documents that were created during the engineering process anyway to create the machine model automatically and in a resource-saving manner.

Other approaches use manually created, company-specific building blocks and rules to create machine models based on module and parameter lists or use a transformation of source code or models of a certain type to create the target model. These M2M transformations are partially supported by additional algorithms, for example, by taking degrees of freedom from 3D CAD models to create behavior models for individual components. The problem often arises that the information from the engineering documents is incomplete and the relationships in the initial models cannot be clearly assigned, so that manual rework is required. The effort of post-processing is about half as high as the total effort would have been without the method [40]. The degree of automation can be increased further, but additional work would be required in the engineering process to create additional documents.

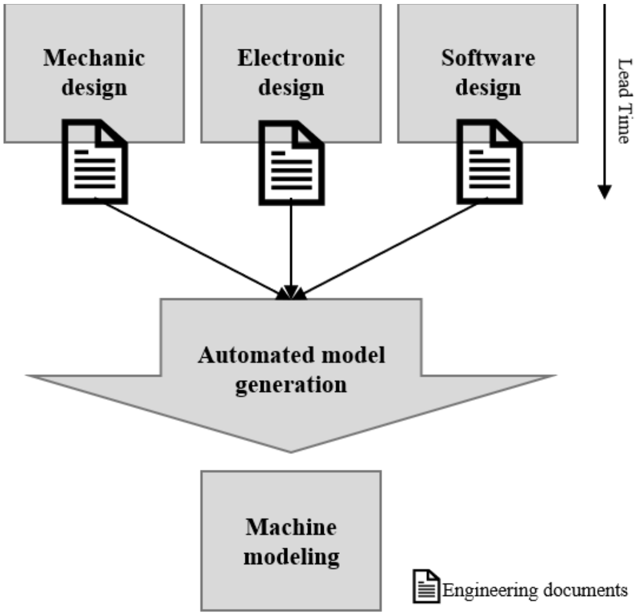


Figure 11: Automated generation of the machine model according to [40].

E. ARIS – people and process view

ARIS, developed by Scheer amongst others in cooperation with the German software company SAP-SE, is an acronym for the architecture of integrated information systems [42]. The underlying model of this approach, which is particularly well-known in Germany, consists of five description views, each with three description levels. The previous form, the so-called ARIS House, is used to reduce complexity and simplify process modeling. The (a), functional view, describes processes and their hierarchical relationships. The (b), organizational view, contains the organizational chart. The (c), data view, contains all company-relevant information objects. The (d), performance view, shows all service, material, and financial services and finally, the (e), process view or control view, integrates all other views (a) to (d) in a time-logical flow chart, such as event-driven process chain (EPC). The description levels are the technical concept, the IT concept, and the implementation level. They serve to represent the business processes for specialists, the implementation of the technical concept in IT-related description models, and the IT-technical realization of the process parts. The software tool ARIS has evolved steadily since its introduction and now consists of several software modules. These enable, among other things, the import of data from data sources such as CRM systems, ERP reports, data warehouses, or Excel tables. In addition, models from UML, MS Visio, BPMN WSDL, XSD, or BPEL can be integrated into the software. Thanks to the uncomplicated import of various file formats and their linking, new information can be implemented quickly. Besides, compatibility with supplier system models can be made easier. Once the system model has been implemented, the ARIS Toolset can be used to automatically create the Quality Management manual, the process and work instructions, job descriptions, the creation of key figures, and process cost accounting.

Furthermore, EIS (Executive Information Systems) takes over the filtering and preparation of decision-relevant information for the management, i.e., data from different sources is merged, and information is offered in a user-friendly way according to different views and levels of aggregation. In addition, data mining techniques are used, which enable the business process owner to navigate in a targeted manner to processes relevant to the investigation. If information objects or attributes are removed from the data model, added to it, or changed, this information automatically leads to an adjustment of the user mask in the system. Automation is also aimed at through the use of object-oriented code generators [43], whereby additional code must be generated manually in some cases and re-delegation takes place in the case of failures caused by the design itself.

F. Which approaches already contribute to reducing the initial and update effort and are they sufficient?

The approaches to partially automated modeling presented here all serve the purpose of reducing the effort involved in creating and updating system models. It will only make sense to use such approaches if this goal can be achieved. The reduction of effort is to be achieved by modeling the five views of eDeCoDe presented above, i.e., only those aspects are considered, which are useful for this purpose. The extent, to which the above-described approaches complement, contradict, or exclude each other as well as eDeCoDe must also be considered. The eDeCoDe views of requirements, components, processes, and, in some cases, the view of the people can be found to some degree in the examined approaches. At least one of the approaches relates to these views, but there is no possibility of partial automation regarding the view of the functions. At the same time, it is noticeable that although each of the approaches is based on a model including its definition, these approaches have little or no overlap.

V. CONCLUSION AND FURTHER WORK

The use of system models is accompanied by many advantages, which are necessary for the success of a company. Especially in the context of the increasing complexity of product and production systems, the system model plays an important role. Therefore, new approaches to system model creation are constantly being published.

In this article, the problem of excessive initial and update efforts in the modeling of production systems was highlighted. It was shown that there are different approaches to depicting production system models and that these contribute to reducing their complexity. However, these approaches have the commonality of manual implementation. Because production systems are made up of numerous system elements and relationships, it is hardly possible for them to be created by individual people. For this reason, the article also critically questioned how far the development of partially automated approaches has progressed. Therefore, approaches of partial automation were also examined, which should reduce the effort of system model creation and updating.

Existing approaches of partial automation of model creation are very branch specific or consider only partial aspects of the overall system, such as the requirements, so that they cannot be used for holistic system description and modeling in a multi-dimensional way, such as eDeCoDe. Other, unspecific approaches to partial automation, on the other hand, do not offer any significant reduction in effort.

The result of this investigation clearly shows that there are approaches that could map individual views of the modeling with, e.g., eDeCoDe in a partially-automated manner. Because these approaches are view-specific, however, the question arises as to whether it is possible to link the view-specific approaches to a holistic approach of partially automated modeling. If this is not the case, it is necessary to develop a new approach to partial automation.

ACKNOWLEDGMENT

The authors thank the German Research Foundation (DFG) for their support of the projects KAUSAL [WI 1234/21-1], ReMaiN [WI 1234/28-1], and FusLa [SCHL 2225/1-1].

REFERENCES

[1] M. Heinrichsmeyer, N. Schlüter, F. Kösling, and A. Ansari, “Validation of a Failure-Cause Searching and Solution-Finding Algorithm in Production based on Complaint Information from the Use Phase,” in The Fifteenth International Conference on Systems, International Journal on Advances in Systems and Measurements IARIA Lisbon, Portugal, pp. 7–12, 2020.

[2] B. S. Onggo, N. Mustafee, A. Smart, A. A. Juan, and O. Molloy, “Symbiotic simulation system: hybrid system model meets big data analytics," 2018 Winter Simulation Conference (WSC), pp. 1358–1369, 2018.

[3] H. Hick, M. Bajzek, and C. Faustmann, “Definition of a system model for model-based development," SN Applied Sciences, vol. 1074, pp. 1074, 10.1007/s42452-019-1069-0, 2019.

[4] A. Canedo, E. Schwarzenbach, and M. A. Al Faruque, “Context-sensitive synthesis of executable functional models of cyber-physical systems,” in Proceedings of the ACM/IEEE 4th International Conference on Cyber-Physical Systems - ICCPS '13, Lu, C.; Kumar, P. R.; Stoleru, R. ACM Press; IEEE New York, New York, USA, pp. 99, 2013.

[5] C. Torens, L. Ebrecht, and K. Lemmer, “Starting Model-Based Testing Based on Existing Test Cases Used for Model Creation,” in 2011 IEEE 11th International Conference on Computer and Information Technology (CIT 2011), IEEE Computer Society IEEE Piscataway, NJ, pp. 320–327, 2011.

[6] R. Haberfellner, O. L. de Weck, E. Fricke, and S. Vössner, “Systems Engineering,” Orell Füssli Verlag; Orell Füssli, vol. 14, Zürich, 2018.

[7] L. Kof, “From Requirements Documents to System Models: A Tool for Interactive Semi-Automatic Translation,” in 2010 18th IEEE International Requirements Engineering Conference (RE 2010), University of Technology, Sydney; IEEE Computer Society; Institute of Electrical and Electronics Engineers; IEEE International Requirements Engineering Conference; RE IEEE Piscataway, NJ, pp. 391–392, 2010.

[8] A.-W. Scheer, “ARIS - Vom Geschäftsprozess zum Anwendungssystem (ARIS - From business process to application system),” Springer, vol. Vierte, durchgesehene Auflage, Berlin, 2002.

[9] B. Stützel, L. Borchardt, T. Illa, and C. Gerling, "*Systems Engineering in Deutschland*," [Online]. Available from: https://www.gfse.de/Dokumente_Mitglieder/se/pubs/downloads/2018-11-06_Prozesswerk_Broschuere_digital.pdf, 01.12.2020.

[10] F. Ehrenmann, “Kosten- und zeiteffizienter Wandel von Produktionssystemen (Cost and time efficient change of production systems),” Springer Gabler, vol. 1, Wiesbaden, 2015.

[11] T. Tomiyama, M. Mäntylä, and S. Finger, “Knowledge Intensive CAD, vol. 1,” Springer US, Boston, MA, s.l., 1996.

[12] M. Heinrichsmeyer, “Entwicklung eines zielgerichteten Fehlerursachensuch- und Lösungsalgorithmus [FusLa] (Development of a targeted error search and solution algorithm [FusLa]),” Bergische Universität Wuppertal, vol. 1, Wuppertal, 2020.

[13] P. Fritzson and P. Bunus, “Modelica - a general object-oriented language for continuous and discrete-event system modeling and simulation,” in Proceedings / 35th Annual Simulation Symposium, SS 2002, IEEE IEEE Computer Society Press Los Alamitos, Calif., pp. 365–380, 2002.

[14] P. Fritzson and V. Engelson, “Modelica — A unified object-oriented language for system modeling and simulation,” Springer, vol. 1, Berlin, 1998.

[15] J. Gausemeier and B. Behmann, “Produkte und Produktionssysteme integrativ konzipieren (Integrative design of products and production systems),” Carl Hanser Fachbuchverlag, vol. 1. Aufl., s.l., 2012.

[16] J. Heihoff-Schwede, C. Bremer, M. Rabe, and C. Tschirner, “Werkzeuge für den Mittelstand–MBSE leicht (Tools for medium scale- MBSE light),” in Tag des Systems Engineering, Schulze, S.-O.; Tschirner, C.; Kaffenberger, R.; Ackva, S. Carl Hanser Fachbuchverlag s.l., pp. 35, 2016.

[17] V. Salehi, G. Florian, and J. Taha, “Implementation of Systems Modeling Language (SysML) in consideration of the CONSENS approach,” in Proceedings of the DESIGN 2018 15th International Design Conference // Design 2018, Marjanović, D.; Štorga, M.; Škec, S.; Bojčetić, N.; Pavković, N. Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia; The Design Society, Glasgow, UK; Fac. of Mechanical Engineering and Naval Architecture Univ Zagreb, pp. 2987–2998, 2018.

[18] M. Maurer and S.-O. Schulze, “Tag des Systems Engineering (Day of Systems Engineering),” Hanser, vol. 1, München, 2015.

[19] S. Friedenthal, “A practical guide to SysML,” Morgan Kaufman, vol. Third edition, Waltham, MA, 2015.

[20] J. Holtman, J. Meyer, W. Schäfer, and U. Nickel, “Eine erweiterte Systemmodellierung zur Entwicklung von softwareintensiven Anwendungen in der Automobilindustrie (An extended system modeling for the development of software intensive applications in the automotive industry),” in Software Engineering 2010 - Workshopband, Engels, G.; Luckey, M. Ges. für Informatik Bonn, pp. 149–158, 2010.

[21] T. Weilkiens, “Systems Engineering with SysML/UML,” Elsevier professional, vol. 1. Aufl., s.l., 2008.

[22] F. Munker, “Ein Ansatz zur anwenderorientierten Systemmodellierung für die interdisziplinäre Produktentwicklung,” KIT; Karlsruhe, vol. 1, Karlsruhe, 2016.

[23] F. Munker and A. Albers, “SystemSketcher – Entstehung eines anwenderorientierten Ansatzes zur interdisziplinären Systemmodellierung (SystemSketcher - Development of a user-oriented approach to interdisciplinary system modeling),” in Tag des Systems Engineering, Schulze, S.-O.; Tschirmer, C. Carl Hanser Fachbuchverlag s.l., pp. 291–300, 2015.

[24] M. Gonia, “Using manufacturing execution systems (MES) to track complex manufacturing processes,” in IEEE/CPMT/SEMI 29th International Electronics Manufacturing Technology Symposium, International Electronics Manufacturing Technology Symposium IEEE Service Center Piscataway, NJ, pp. 171–173, 2004.

[25] J. Kletti, “Manufacturing Execution Systems - MES,” Springer-Verlag Berlin Heidelberg, vol. 1, Berlin, Heidelberg, 2007.

[26] P. Winzer, “Generic Systems Engineering,” Springer Vieweg Verlag, vol. 2, Berlin, Heidelberg, 2016.

[27] J.-P. G. Nicklas, “Ansatz für ein modellbasiertes Anforderungsmanagement für Unternehmensnetzwerke (Approach for a model-based requirements management for enterprise networks),” Shaker Verlag, vol. 1, Aachen, 2016.

[28] S. Schlund and P. Winzer, “DeCoDe-Modell zur anforderungsgerechten Produktentwicklung (DeCoDe model for requirements-based product development),” in "Das ist gar kein Modell!" (It is not a model), Bandow, G.; Holzmüller, H. H. Gabler Wiesbaden, pp. 277–293, 2010.

[29] A. Hahn, S. Häusler, and S. Große Austing, “Quantitatives Entwicklungsmanagement (Quantitative development management),” Springer Vieweg, vol. 1, Berlin, 2013.

[30] O. Bielefeld, H. Dransfeld, N. Schlüter, and P. Winzer, “Development of an Innovative Approach for Complex, Causally Determined Failure Chains," Management and Production Engineering Review, vol. 3, pp. 3–12, 10.1515/mper-2017-0023, 2017.

[31] D. Krause and N. Luhmann, “Luhmann-Lexikon,” Stuttgart: UTB; Enke, vol. 2., vollst. überarb., erw. und aktualisierte Aufl., Stuttgart, 2005 // 1999.

[32] O. Bielefeld, N. Schlüter, and P. Winzer, “Development of a methodological approach for requirements management in cross-company networks (ReMaiN),” in Proceedings M2D2017, Silva Gomes, J. F.; Meguid, S. A. Edições INEGI Porto, pp. 1109–1124, 2017.

[33] S. Marchlewitz, J.-P. Nicklas, and P. Winzer, “Using system engineering for improving autonomous robot performance,” in 2015 10th System of Systems Engineering Conference (SoSE 2015), Institute of Electrical and Electronics Engineers; IEEE Systems, Man, and Cybernetics Society; IEEE Reliability Society; International Council on Systems Engineering; System of Systems Engineering Conference; International IEEE Conference on Systems of Systems Engineering; SoSE IEEE Piscataway, NJ, pp. 65–70, 2015.

[34] A. Götz and C. Donges, “Automatisierter Übergang vom dokumenten- zum modell-zentrierten Requirements Engineering als Ausgangsbasis für MBSE (Automated transition from document to model-centric requirements engineering as a starting point for MBSE),” in Tag des Systems Engineering, Schulze, S.-O.; Tschirner, C.; Kaffenberger, R.; Ackva, S. Carl Hanser Verlag GmbH & Co. KG München, pp. 301–310, 2017.

[35] El-Haik. Basem and K. Yang, “The components of complexity in engineering design,” in IEEE Transaction, IEEE Springer Berlin, pp. 925–934, 10.1999.

[36] A. Burger, S. Reiter, A. Viehl, O. Bringmann, and W. Rosenstiel, "*Systemmodellierung zur Fehlereffektsimulation (System modeling for failure effect simulation)*," [Online]. Available from: https://www.researchgate.net/publica-tion/273004392_Systemmodellierung_zur_Fehlereffektsimulation, 01.12.2020.

[37] D. Krenczyk, “Automatic Generation Method of Simulation Model for Production Planning and Simulation Systems Integration," Advanced Materials Research, pp. 825–829, 10.4028/www.scientific.net/AMR.1036.825, 2014.

[38] G. Rempp, M. Akermann, M. Löffler, J. Lehmann, and M. Starzmann, “Model Driven SOA,” Springer, vol. 1, Berlin, 2011.

[39] P. Bocciarelli, A. D'Ambrogio, A. Mastromattei, and A. Giglio, “Automated development of web-based modeling services for MSaaS platforms,” in International Symposium on Model-driven Approaches for Simulation Engineering

(Mod4Sim'17), D'Ambrogio, A.; Durak, U.; Çetinkaya, D. Curran Associates Inc Red Hook, NY, pp. 1–12, 2017.

[40] A. Kufner, “Automatisierte Erstellung von Maschinenmodellen für die Hardware-in-the-Loop-Simulation von Montagemaschinen (Automated creation of machine models for hardware-in-the-loop simulation of assembly machines),” Jost-Jetter, vol. 1, Heimsheim, 2012.

[41] R. Angerbauer, R. Buck, U. Doll, M. Hackel, K.-H. Kayser, M. Klebl, H. Mack, R. Siegler, F. Wascher, and R. Würslin, “Aquimo,” VDMA-Verl., vol. 1, Frankfurt, M., 2010.

[42] A.-W. Scheer, “ARIS - Business Process Modeling,” Springer Berlin Heidelberg, vol. Third Edition, Berlin, Heidelberg, s.l., 2000.

[43] A.-W. Scheer, “ARIS - Modellierungsmethoden, Metamodelle, Anwendungen (ARIS - Modeling methods, metamodels, applications),” Springer Berlin Heidelberg, vol. Dritte, völlig neubearbeitete und erweiterte Auflage, Berlin, Heidelberg, 1998.

Point Cloud Mapping and Merging in GNSS-Denied and Dynamic Environments Using Onboard Scanning LiDAR

Seiya Tanaka, Chisato Koshiro, Misato Yamaji
Graduate School of Science and Engineering
Doshisha University
Kyotanabe, Kyoto 610-0394 Japan
e-mail: {ctwd0144, ctwf0118, ctwf0148}@mail4.doshisha.ac.jp

Masafumi Hashimoto, Kazuhiko Takahashi
Faculty of Science and Engineering
Doshisha University
Kyotanabe, Kyoto 610-0394 Japan
e-mail: {mhashimo, katakaha}@mail.doshisha.ac.jp

Abstract— **This paper presents a 3D point cloud mapping and merging in Global Navigation Satellite Systems (GNSS)-denied and dynamic environments using only a scanning Light Detection And Ranging (LiDAR) mounted on a vehicle. Distortion in scan data from the LiDAR is corrected by estimating the vehicle's pose (3D positions and attitude angles) in a period shorter than the LiDAR scan period using Normal Distributions Transform (NDT) scan matching and Extended Kalman Filter (EKF). The corrected scan data are mapped onto an elevation map. Static and moving scan data, which originate from static and moving objects, respectively, in the environments are classified using the occupancy grid method. Only the static scan data are utilized to generate several submaps in different small areas using NDT-based Simultaneous Localization And Mapping (NDT SLAM) and Graph SLAM. These submaps are merged using Graph SLAM. Experimental results obtained in outdoor residential and urban road environments show the LiDAR-based mapping and merging via EKF and NDT-Graph SLAM provide accurate maps in GNSS-denied and dynamic environments.**

Keywords-LiDAR; point cloud map; mapping and merging; NDT-Graph SLAM.

I. Introduction

This paper is an extended and improved version of an earlier paper presented at the IARIA Conference on Systems (ICONS 2020) [1] in Lisbon.

Recently, many studies have been conducted on the autonomous driving and active safety of vehicles, such as automobiles and personal mobility vehicles, and on autonomous robots for last-mile and first-mile automation. Important technologies from these studies include environmental map generation (mapping) [2] and map-matching-based self-pose estimation by vehicles using generated maps [3]. Many related studies used cameras, radars, and Light Detection And Ranging (LiDAR) [4][5].

In this paper, we focus on mapping with a scanning LiDAR mounted on a vehicle. Compared with camera-based mapping, LiDAR-based mapping is robust to lighting conditions and requires less computational time. Furthermore, the accuracy of LiDAR-based mapping is better than that of radar-based mapping due to the higher spatial resolution of LiDAR. For these reasons, we focus on LiDAR-based mapping.

In Intelligent Transportation Systems (ITS) domains, mobile mapping systems are used for mapping in wide road environments, such as highways and motorways [6]. We studied a method for point cloud mapping in narrow road environments, such as residential roads in urban and mountainous environments, using only a vehicle-mounted LiDAR [7]. The generated map could be applied to the autonomous driving and navigation of various smart vehicles, such as intelligent wheelchairs, personal mobility devices, and delivery robots [8]. The generated maps may also be utilized in various social services, such as disaster prevention and mitigation.

Although mapping systems often utilize position information from Global Navigation Satellite Systems (GNSS) [9], the accuracy of GNSS positioning is decreased in urban and mountainous areas due to the blockage, reflection, and diffraction caused by buildings and mountains. In addition, mapping systems designed for mapping in static environments generate inconsistent maps in practical dynamic environments that have moving objects, such as cars and pedestrians.

To address these problems, many studies have been conducted on LiDAR-based Simultaneous Localization And Mapping (SLAM) [9]. However, LiDAR-based SLAM in GNSS-denied and dynamic environments, such as urban street canyons in which the GNSS accuracy deteriorates and vehicles and people move, remains a significant challenge. This paper presents a point cloud mapping that uses only an onboard scanning LiDAR in GNSS-denied and dynamic environments. To do so, this technique integrates three methods that we previously proposed: distortion correction of the LiDAR scan data [10], extraction of scan data related to static objects from the entire LiDAR scan data [11][12], and point cloud mapping based on Normal Distributions Transform (NDT) and Graph-based SLAM [7]. The mapping performance by the proposed method is shown through experimental results in outdoor road environments.

The rest of this paper is organized as follows. Section II presents an overview of related work, and Section III describes the experimental system. Section IV explains the correction method of LiDAR scan data distortion, and Section V presents the extraction method of static scan data, which are related to static objects (removal of moving scan data, which are related to moving objects) from the entire LiDAR scan data. Section VI describes the mapping and merging methods based on NDT and Graph SLAM (called NDT-Graph SLAM). Finally, Section VII explains the experiments conducted to show the performance of our method, followed by the conclusions in Section VIII.

II. Related Work

The main contribution of this paper is the conduct of LiDAR-based SLAM in GNSS-denied and dynamic environments by integrating components that we previously proposed: distortion correction of the LiDAR scan data [10], extraction of the scan data related to static objects from the entire LiDAR scan data [11][12], and point cloud mapping and merging based on NDT-Graph SLAM [7].

LiDAR-based SLAM is performed by mapping LiDAR scan data captured in a sensor coordinate frame onto a world coordinate frame using the vehicle's self-pose (position and attitude angle) information. The LiDAR obtains range measurements by scanning LiDAR beams. Thus, when the vehicle moves, the entire scan data within one scan (LiDAR beam rotation of 360° in a horizontal plane) cannot be obtained at the same pose of the vehicle. Therefore, if the entire scan data obtained within one scan are mapped onto the world coordinate frame using information about the vehicle's pose at a single point in time, distortion will arise in mapping. This distortion can be corrected by determining the vehicle's pose more frequently than the LiDAR scan period, i.e., for every LiDAR measurement in the scan.

Many distortion correction methods have been proposed [13][14][15]. However, most methods used additional sensors, such as odometer, Inertial Measurement Unit (IMU), and GNSS. Simple interpolation algorithms were also applied to determine a vehicle's pose more frequently than the LiDAR scan period. Unlike conventional methods, we corrected the distortion of LiDAR scan data using only the LiDAR information via Extended Kalman Filter (EKF) [10]. Our distortion correction method performed well.

When environmental features such as planes and pole-like objects are available, scan matching (such as NDT [16] and Iterative Closest Points (ICP) [17] methods) is applied to LiDAR-based SLAM in GNSS-denied environments [18]. Scan matching is adopted to calculate the transformation between LiDAR scans. The LiDAR-based SLAM is then performed based on the calculated continuous transformation. One of cons in the LiDAR-based SLAM is the drift (degradation of the accuracy over time) due to the accumulation error. To reduce the drift, Graph SLAM [19] is employed in conjunction with LiDAR-based SLAM. Another effective approach toward reducing the drift by LiDAR-based SLAM is submap generation and merging; the drift can be avoided by allowing short trajectories per submap [20][21].

We presented a mapping method in GNNS denied environments based on NDT-Graph SLAM [7]. A vehicle equipped with a LiDAR was moved such that loops could be made in road networks, and several submaps (maps of different small areas) were generated using NDT-Graph SLAM. Several submaps were also merged using Graph SLAM. Such approach to submap generation and merging makes it easy to update and maintain maps. However, further improvement is needed in the accuracy of submap merging. In addition, since a static world was assumed in our previous work, the presence of moving objects in practical dynamic environments deteriorates mapping performance. Then, improvements are required in the mapping method in dynamic environments.

In dynamic environments, LiDAR scan data can be classified into two types, namely, scan data originating from moving objects (moving scan data), and those originating from static objects (static scan data), such as buildings, trees, and traffic poles. For accurate mapping, the moving scan data have to be removed; only the static scan data will be utilized. This problem is addressed by SLAM-Moving Object Tracking (MOT) or SLAM-Detection And Tracking of Moving Objects (DATMO) approaches [22][23].

Apart from mapping, we have studied MOT and DATMO in crowded dynamic environments [11][12] for driving safety. Our moving-object detection method in MOT and DATMO was based on the occupancy grid method, which used the cell occupancy time and is simpler than usual probabilistic occupancy grid methods [24]. Our moving-object detection method will accurately remove moving scan data from the entire LiDAR scan data captured in dynamic environments and generate static maps.

III. Experimental System

As shown in Figure 1, our small experimental vehicle is equipped with a 32-layer scanning LiDAR (Velodyne HDL-32E). The maximum range of the LiDAR is 70 m, the horizontal viewing angle is 360° with a resolution of 0.16°, and the vertical viewing angle is 41.34° with a resolution of 1.33°. The LiDAR provides 384 measurements (the object's 3D position and reflection intensity) every 0.55 ms (at 2° horizontal angle increments). The period for the LiDAR beam to complete one rotation (360°) in the horizontal direction is 100 ms, and 70,000 measurements are obtained in one rotation.

In this paper, one rotation of the LiDAR beam in the horizontal direction (360°) is considered one scan, and the data obtained from this scan are called scan data. Moreover, the LiDAR scan period (100 ms) is denoted as τ and the

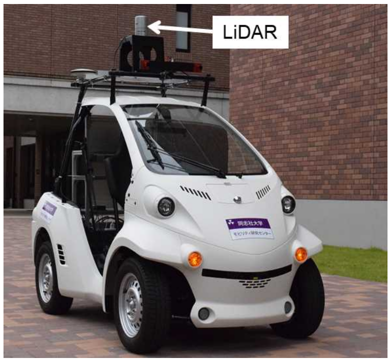


Figure 1. Overview of experimental vehicle.

scan-data observation period (0.55 ms) as $\Delta\tau$.

To evaluate the SLAM performance, the vehicle is equipped with a GNSS/Inertial Navigation System (INS) unit (Novatel SPAN-CPT). The GNSS/INS unit outputs the vehicle's 3D position and attitude angle (roll, pitch, and yaw angles) every 100 ms. The horizontal and vertical position errors (Root Mean Square, RMS) are 0.02 m and 0.03 m, respectively. The roll and pitch angle errors (RMS) are both 0.02°, and the yaw angle error (RMS) is 0.06°.

IV. Distortion Correction of LiDAR Scan Data

This section describes the mapping method of scan data using NDT scan matching and distortion correction of LiDAR scan data using EKF.

A. NDT Scan Matching

The vehicle coordinate frame Σ_b (O_b-$x_b y_b z_b$) is defined in Figure 2. The origin O_b is the center of the rear wheel axle of the vehicle; the x_b, y_b, and z_b axes are the heading direction, the direction of the rear wheel axle, and the direction toward the sky, respectively. Although the LiDAR scan data are captured by the sensor coordinate frame fixed at the LiDAR, the objects' 3D positions in the scan data are always transformed to those in Σ_b . For convenience, the scan data are hereafter assumed to be captured in Σ_b .

When LiDAR scan data are captured in one scan, the scan data related to road surfaces are first removed using a method described in Section V, and the scan data related to objects are mapped onto a 3D grid map (voxel map) represented in Σ_b . A voxel grid filter [25] is applied to downsize the scan data. The block used for the voxel grid filter is a cube with a side length of 0.2 m.

A local coordinate frame Σ_W (O_W -$x_W y_W z_W$) is defined in Figure 2. Σ_W coincides with Σ_b when the vehicle starts to generate the submap. In Σ_w , a voxel map with a voxel size of 1 m is used for NDT scan matching. For the i-th (i = 1, 2, …n) measurement in the scan data, the position vector in Σ_b is denoted as $\boldsymbol{p}_{bi}$ and that in Σ_w as $\boldsymbol{p}_i$. The following relation is then given:

$$\begin{pmatrix} \boldsymbol{p}_i \\ 1 \end{pmatrix} = \boldsymbol{T}(\boldsymbol{x}) \begin{pmatrix} \boldsymbol{p}_{bi} \\ 1 \end{pmatrix} \tag{1}$$

where $\boldsymbol{x} = (x, y, z, \phi, \theta, \psi)^T$ is the vehicle's pose. $(x, y, z)^T$ and $(\phi, \theta, \psi)^T$ are the 3D position and attitude angle (roll, pitch, and yaw angles) of the vehicle, respectively, in Σ_W . $\boldsymbol{T}(\boldsymbol{x})$ is the following homogeneous transformation matrix:

$$T(x) = \begin{pmatrix} \cos\theta\cos\psi & \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & x \\ \cos\theta\sin\psi & \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi & y \\ -\sin\theta & \sin\phi\cos\theta & \cos\phi\cos\theta & z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The scan data obtained at the current time $t\tau$ (t = 0, 1, 2, …), $\boldsymbol{p}_{b(t)} = \{\boldsymbol{p}_{b1(t)}, \boldsymbol{p}_{b2(t)}, \cdots\}$, are called the new input scan, and the scan data obtained in the previous time, i.e., before $(t-1)\tau$, $\boldsymbol{P} = \{\boldsymbol{P}_{(0)}, \boldsymbol{P}_{(1)}, \cdots, \boldsymbol{P}_{(t-1)}\}$, are called the reference scan (environmental map).

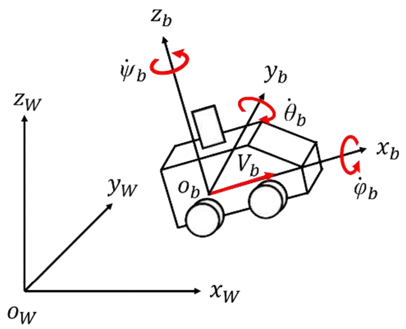


Figure 2. Notation related to vehicle motion.

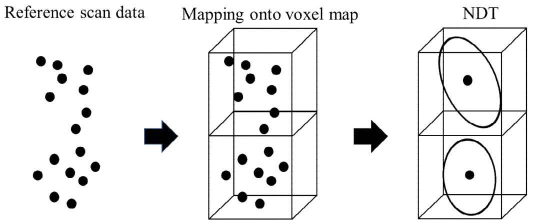


Figure 3. Normal distributions transform of reference scan data.

NDT scan matching [16] conducts a normal distribution transformation for the reference scan data in each grid on a voxel map. It calculates the mean and covariance of the LiDAR measurement positions, as shown in Figure 3. The vehicle's pose $\boldsymbol{x}_{(t)}$ at $t\tau$ is determined by matching the new input scan at $t\tau$ with the reference scan data obtained prior to $(t-1)\tau$. The vehicle's pose can be calculated by maximizing the following likelihood function:

$$\Lambda = \prod_{i=1}^{n} \exp\left(-\frac{1}{2} (\boldsymbol{p}_i(t) - \boldsymbol{q}_i)^T \boldsymbol{\Omega}_i^{-1} (\boldsymbol{p}_i(t) - \boldsymbol{q}_i) \right) \tag{2}$$

where $\boldsymbol{q}_i$ and $\boldsymbol{\Omega}_i$ are the mean and covariance, respectively, of the reference scan in the i-th voxel. $\boldsymbol{p}_i$ is the new input scan in the i-th voxel.

The vehicle's pose is used for conducting a coordinate transform with (1). The new input scan can then be mapped to Σ_W , and the reference scan is updated. The downsized scan data are only used to calculate the vehicle's pose using NDT scan matching for a small computational cost.

In this study, we use the Point Cloud Library (PCL) [26] for NDT scan matching.

B. Distortion Correction of LiDAR Scan Data

A motion model of the vehicle is first described for the EKF-based correction of LiDAR scan data distortion.

As shown in Figure 2, the vehicle's linear velocity in Σ_b is defined as V_b (the velocity in the x_b-axis direction), and the angular velocities about the x_b, y_b, and z_b axes are defined as $\dot{\phi}_b$, $\dot{\theta}_b$, and $\dot{\psi}_b$, respectively. If the vehicle is assumed to move at nearly constant linear and angular velocities, the following motion model can then be derived:

$$\begin{pmatrix} x_{(t+1)} \\ y_{(t+1)} \\ z_{(t+1)} \\ \phi_{(t+1)} \\ \theta_{(t+1)} \\ \psi_{(t+1)} \\ V_{b(t+1)} \\ \dot{\phi}_{b(t+1)} \\ \dot{\theta}_{b(t+1)} \\ \dot{\psi}_{b(t+1)} \end{pmatrix} = \begin{pmatrix} x_{(t)} + a_{1(t)}\cos\theta_{(t)}\cos\psi_{(t)} \\ y_{(t)} + a_{1(t)}\cos\theta_{(t)}\sin\psi_{(t)} \\ z_{(t)} - a_{1(t)}\sin\theta_{(t)} \\ \phi(t) + a_2(t) + \{a_3(t)\sin\phi(t) + a_4(t)\cos\phi(t)\}\tan\theta(t) \\ \theta(t) + \{a_3(t)\cos\phi(t) - a_4(t)\sin\phi(t)\} \\ \psi(t) + \{a_3(t)\sin\phi(t) + a_4(t)\cos\phi(t)\}\dfrac{1}{\cos\theta(t)} \\ V_{b(t)} + \tau w_{\dot{V}_b} \\ \dot{\phi}_{b(t)} + \tau w_{\ddot{\phi}_b} \\ \dot{\theta}_{b(t)} + \tau w_{\ddot{\theta}_b} \\ \dot{\psi}_{b(t)} + \tau w_{\ddot{\psi}_b} \end{pmatrix} \tag{3}$$

where $a_1 = V_b\tau + \tau^2 w_{\dot{V}_b}/2$, $a_2 = \dot{\phi}_b\tau + \tau^2 w_{\ddot{\phi}_b}/2$, $a_3 = \dot{\theta}_b\tau + \tau^2 w_{\ddot{\theta}_b}/2$, and $a_4 = \dot{\psi}_b\tau + \tau^2 w_{\ddot{\psi}_b}/2$. $w_{\dot{V}_b}$, $w_{\ddot{\phi}_b}$, $w_{\ddot{\theta}_b}$, and $w_{\ddot{\psi}_b}$ are the acceleration disturbances.

Equation (3) is expressed in vector form as follows:

$$\boldsymbol{\xi}_{(t+1)} = \boldsymbol{f}[\boldsymbol{\xi}_{(t)}, \boldsymbol{w}, \tau] \tag{4}$$

where $\boldsymbol{\xi} = (x, y, z, \phi, \theta, \psi, V_b, \dot{\phi}_b, \dot{\theta}_b, \dot{\psi}_b)^T$ and $\boldsymbol{w} = (w_{\dot{V}_b}, w_{\ddot{\varphi}_b}, w_{\ddot{\theta}_b}, w_{\ddot{\psi}_b})^T$.

The vehicle's pose obtained at $t\tau$ using NDT scan matching is defined as $\boldsymbol{z}_{NDT(t)} (= \hat{\boldsymbol{x}}_{(t)})$. The measurement equation is then

$$\boldsymbol{z}_{NDT(t)} = \boldsymbol{H}\boldsymbol{\xi}_{(t)} + \Delta\boldsymbol{z}_{NDT(t)} \tag{5}$$

where $\Delta\boldsymbol{z}_{NDT}$ is the measurement noise, and $\boldsymbol{H}$ is the following measurement matrix:

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The correction flow of LiDAR scan data is shown in Figure 4. The LiDAR scan period τ is 100 ms, and the scan-data observation period $\Delta\tau$ is 0.55 ms. When the scan data are mapped onto Σ_W using the vehicle's pose, which is calculated every LiDAR scan period, distortion arises in the environmental map. This distortion of the LiDAR scan data is therefore corrected by estimating the vehicle's pose using the EKF every scan-data observation period $\Delta\tau$.

The state estimate and its error covariance obtained at $(t-1)\tau$ using the EKF are denoted as $\hat{\boldsymbol{\xi}}_{(t-1)}$ and $\boldsymbol{\Gamma}_{(t-1)}$, respectively. From these quantities, the EKF gives the state prediction $\hat{\boldsymbol{\xi}}_{(t-1,1)}$ and its error covariance $\boldsymbol{\Gamma}_{(t-1,1)}$ at $(t-1)\tau + \Delta\tau$ as follows:

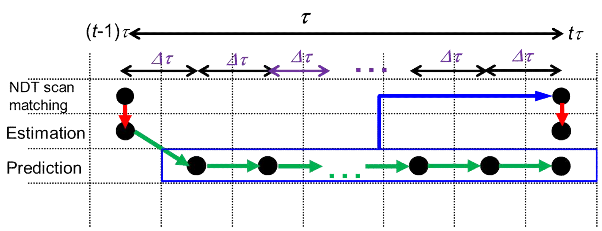


Figure 4. Flow of distortion correction.

$$\left.\begin{aligned} \hat{\boldsymbol{\xi}}_{(t-1,1)} &= \boldsymbol{f}[\hat{\boldsymbol{\xi}}_{(t-1)}, 0, \Delta\tau] \\ \boldsymbol{\Gamma}_{(t-1,1)} &= \boldsymbol{F}_{(t-1)}\boldsymbol{\Gamma}_{(t-1)}\boldsymbol{F}_{(t-1)}^T + \boldsymbol{G}_{(t-1)}\boldsymbol{Q}\boldsymbol{G}_{(t-1)}^T \end{aligned}\right\} \tag{6}$$

where $\boldsymbol{F} = \partial\boldsymbol{f}/\partial\hat{\boldsymbol{\xi}}$ and $\boldsymbol{G} = \partial\boldsymbol{f}/\partial\boldsymbol{w}$. $\boldsymbol{Q}$ is the covariance matrix of the plant noise $\boldsymbol{w}$.

By a similar calculation, the state prediction $\hat{\boldsymbol{\xi}}_{(t-1,j)}$ and its error covariance $\boldsymbol{\Gamma}_{(t-1,j)}$ at $(t-1)\tau + j\Delta\tau$ (where j = 1, 2, …,180) can be obtained by

$$\left.\begin{aligned} \hat{\boldsymbol{\xi}}_{(t-1,j)} &= \boldsymbol{f}[\hat{\boldsymbol{\xi}}_{(t-1,j-1}, 0, \Delta\tau] \\ \boldsymbol{\Gamma}_{(t-1,j)} &= \boldsymbol{F}_{(t-1,j-1)}\boldsymbol{\Gamma}_{(t-1,j-1)}\boldsymbol{F}_{(t-1,j-1)}^T \\ &\quad + \boldsymbol{G}_{(t-1,j-1)}\boldsymbol{Q}\boldsymbol{G}_{(t-1,j-1)}^T \end{aligned}\right\} \tag{7}$$

In the state prediction $\hat{\boldsymbol{\xi}}_{(t-1,j)}$, the elements related to the vehicle's pose $(x, y, z, \phi, \theta, \psi)$ are denoted as $\hat{\boldsymbol{X}}_{(t-1,j)}$. Using (1) and the pose prediction, the scan data $\boldsymbol{p}_{bi(t-1,j)}$ in Σ_b obtained at $(t-1)\tau + j\Delta\tau$ can be transformed to $\boldsymbol{p}_{i(t-1,j)}$ in Σ_W as follows:

$$\begin{pmatrix} \boldsymbol{p}_{i(t-1,j)} \\ 1 \end{pmatrix} = \boldsymbol{T}(\hat{\boldsymbol{X}}_{(t-1,j)}) \begin{pmatrix} \boldsymbol{p}_{bi(t-1,j)} \\ 1 \end{pmatrix} \tag{8}$$

Since the LiDAR scan period τ is 100 ms, and the scan-data observation period $\Delta\tau$ is 0.55 ms, the time $t\tau$ is equal to $(t-1)\tau + 180\Delta\tau$. Using the pose prediction $\hat{\boldsymbol{X}}_{(t-1,180)}$ at $t\tau$, the scan data $\boldsymbol{p}_{i(t-1,j)}$ at $(t-1)\tau + j\Delta\tau$ in Σ_W are transformed into the scan data $\boldsymbol{p}^*_{bi(t)}$ at $t\tau$ in Σ_b as follows:

$$\begin{pmatrix} \boldsymbol{p}^*_{bi(t)} \\ 1 \end{pmatrix} = T(\hat{\boldsymbol{X}}_{(t-1,180)})^{-1} \begin{pmatrix} \boldsymbol{p}_{i(t-1,j)} \\ 1 \end{pmatrix} \tag{9}$$

Using the corrected scan data $\boldsymbol{p}^*_{b(t)} = \{\boldsymbol{p}^*_{b1(t)}, \boldsymbol{p}^*_{b2(t)}, \cdots\}$ within one scan (LiDAR beam rotation of 360° in a horizontal plane) as the new input scan, NDT scan matching can accurately calculate the vehicle's pose $\boldsymbol{z}_{NDT(t)}$ at $t\tau$. Based on (4) and (5), the EKF then gives the state estimate $\hat{\boldsymbol{\xi}}_{(t)}$ and its error covariance $\boldsymbol{\Gamma}_{(t)}$ at $t\tau$ by

$$\left.\begin{aligned} \hat{\boldsymbol{\xi}}_{(t)} &= \hat{\boldsymbol{\xi}}_{(t-1,180)} + \boldsymbol{K}_{(t)}\{\boldsymbol{z}_{NDT}{}_{(t)} - \boldsymbol{H}\hat{\boldsymbol{\xi}}_{(t-1,180)}\} \\ \boldsymbol{\Gamma}_{(t)} &= \boldsymbol{\Gamma}_{(t-1,180)} - \boldsymbol{K}_{(t)}\boldsymbol{H}\boldsymbol{\Gamma}_{(t-1,180)} \end{aligned}\right\} \qquad (10)$$

where $\hat{\boldsymbol{\xi}}_{(t-1,180)}$ and $\boldsymbol{\Gamma}_{(t-1,180)}$ are the state prediction and its error covariance at $t\tau$ $(=(t-1)\tau+180\Delta\tau)$ respectively. $\boldsymbol{K}_{(t)} = \boldsymbol{\Gamma}_{(t-1,180)}\boldsymbol{H}^T \ \boldsymbol{S}^{-1}{}_{(t)}$ and $\boldsymbol{S}_{(t)} = \boldsymbol{H}\boldsymbol{\Gamma}_{(t-1,180)}\boldsymbol{H}^T + \boldsymbol{R}$. $\boldsymbol{R}$ is the covariance matrix of the measurement noise $\Delta\boldsymbol{z}_{NDT}$.

The corrected scan data $\boldsymbol{P}_b^*{}_{(t)}$ are mapped onto Σ_W using the pose estimate calculated by (10), and the distortion in the environmental map can then be removed.

V. Extraction of Static Scan Data

In dynamic environments, which have moving objects, such as cars, two-wheelers, and pedestrians, LiDAR scan data related to moving objects (moving scan data) have to be removed from the entire scan data, and only scan data related to static objects (static scan data), such as buildings and trees, have to be utilized in mapping.

In the extraction of static scan data, the LiDAR scan data are classified into two types, namely, scan data originating from road surfaces (road-surface scan data) and those originating from objects (object scan data), based on the following rule-based method.

As shown in Figure 5, 32 measurements captured every horizontal resolution (0.16°) of the LiDAR are considered. The measurement r_1, which is the closest measurement to the LiDAR, is assumed to be the measurement belonging to road surfaces. We obtain the angle of a line connecting the adjacent measurements r_1 and r_2 relative to the *xy*-plane in Σ_W. If the angle is less than 15°, the measurement r_2 is determined to belong to road surfaces. If it is larger than 15°, the measurement r_2 is determined to belong to objects. By repeating this process for all LiDAR scan data, we can distinguish the scan data related to objects (blue points in Figure 5) and those related to road surfaces (red points). If the threshold for discriminating the scan data related to road surfaces and objects is small, slopes is mis-detected as objects. In general, the steep slope of vehicles is less than about 6°. The threshold is therefore set to 15°.

The object scan data are mapped onto an elevation map represented in Σ_W. In this paper, the cell of the elevation map is a square with a side length of 0.3 m. The height of each cell is the maximum height of multiple scan data mapped onto the cell.

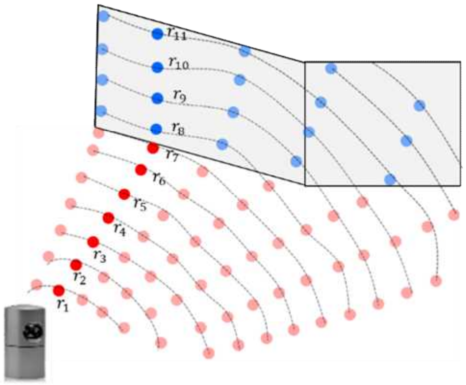


Figure 5. Extraction of LiDAR scan data related to objects.

A cell containing scan data is called an occupied cell. For the moving scan data, the time to occupy the same cell is short (less than 0.8 s in this paper), whereas for the static scan data, the time is long (not less than 0.8 s). Therefore, using the occupancy grid method, which is based on the cell occupancy time [11][12], the occupied cells are classified into two types of cells, namely, moving and static cells, which are occupied by the moving and static scan data, respectively. Cells that the LiDAR cannot identify because of obstructions are defined as unknown cells, and their cell occupancy time is not counted.

Since the scan data related to an object usually occupy multiple cells, adjacent occupied cells with almost the same height are clustered. In general, moving and static cells coexist in the same clustered cells. If the number of moving cells in clustered cells is not less than a threshold *TH*, these clustered cells are then decided as the moving-cell group; otherwise as the static-cell group. *TH* is given by the following sigmoid function:

$$TH = 0.5 + \frac{0.2}{1+\exp(5-0.3s)} \qquad (11)$$

where s is the number of cells that constitute the cell group.

The above equation means that the threshold is dynamically determined to be 50 %–70 % according to the number of cells s. In our experience, since the speed of small (large) moving objects, such as pedestrians (cars), is low (high), the number of moving cells belonging to a cell group is small (large). To improve the performance of the moving-object detection, the threshold is set to 50 % (70 %) for small (large) objects with a small (large) number of occupied cells. The scan data in clustered static cells are applied to mapping.

When moving objects pause, such as vehicles pausing at red lights, the occupancy grid-based method often misidentifies their scan data as static scan data. To address this problem, road-surface scan data are mapped onto the elevation map, and the cells where the road-surface scan data have been occupied for several scans are determined as road-surface cells. If the road-surface cells contain object scan data, these data are always determined as moving scan data and removed from the entire scan data.

VI. Submap Generation and Merging

This section describes the methods of submap generation and merging based on NDT-Graph SLAM. For a clear explanation, we consider the generation and merging submaps 1 and 2, which are shown in Figure 6.

A. Submap Generation

In each submap, a local coordinate frame Σ_{Wi} (O_{Wi} -x_{Wi} y_{Wi} z_{Wi}) is defined, where i = 1, 2; Σ_{Wi} coincides with Σ_b when the vehicle starts to generate the submap i.

The vehicle's poses are mapped onto a factor graph (pose graph), as shown in Figure 7. In this figure, the vehicle's poses are represented as the graph nodes (black triangles), and the relative poses between two neighboring nodes are represented as the graph edges (black arrows). The vehicle's poses are calculated by NDT SLAM every 100 ms (LiDAR scan period).

To recognize whether or not the vehicle has already visited a place (called revisit node or loop), the candidate of the revisit nodes is first obtained using the self-location information of the vehicle, which is estimated by NDT SLAM. If the distance of an old node from the current node is smaller than 10 m, as shown in Figure 8, the old node is recognized as a candidate of the revisit nodes.

Thereafter, the Loop Probability Indicator (LPI) [27] is calculated using LiDAR scan data captured at the candidate of the revisit and current nodes. Each grid of the voxel map is first classified into three types of voxels: line, plane, and the other voxels in Figure 9. Three eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$) are calculated from LiDAR scan data in voxels based on the principal component analysis. When λ_2 / λ_1 is no more than 0.1, the voxel is decided as being of line type (Figure 9 (a)); when λ_3 / λ_2 is no more than 0.1, the voxel is decided as being of plane type (Figure 9 (b)); when λ_2 / λ_1 and λ_3 / λ_2 are more than 0.1, the voxel is decided as being of another type (Figure 9 (c)).

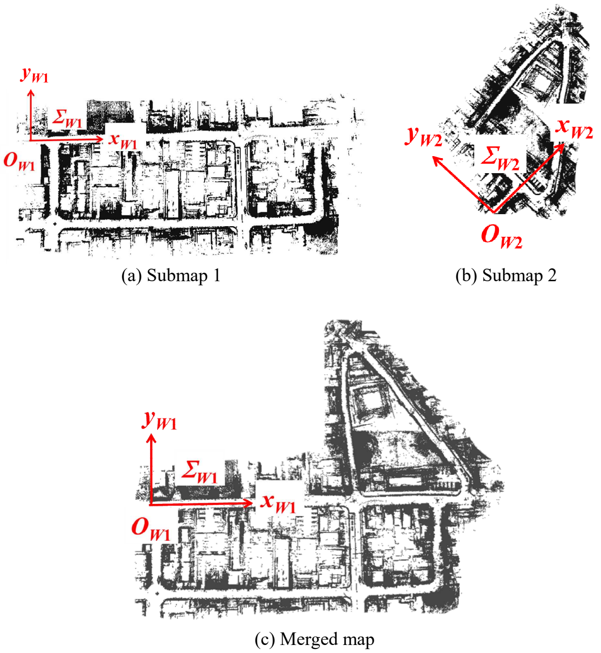
Based on the surface normal vector of the plane voxels in Σ_b, these plane voxels are further divided into nine classes: (1, 0, 0), (0, 1, 0), (0, 0, 1), $(1/\sqrt{2},1/\sqrt{2},0)$, $(1/\sqrt{2},-1/\sqrt{2},0)$, $(1/\sqrt{2},0,\ 1/\sqrt{2})$, $(-1/\sqrt{2},0,1/\sqrt{2})$, $(0,1/\sqrt{2},1/\sqrt{2})$, and $(0,-1/\sqrt{2},1/\sqrt{2})$.

Two feature descriptors $\boldsymbol{U} = (u_1, u_2, \cdots, u_{11})^T$ and $\boldsymbol{V} = (v_1, v_2, \cdots, v_{11})^T$ are defined. $\boldsymbol{U}$ is calculated from LiDAR scan data captured at the candidate of the revisit nodes, and $\boldsymbol{V}$ is calculated from the LiDAR scan data at the current node. u_1 and v_1 are the numbers of line voxels in the voxel map. $u_2 - u_{10}$ and $v_2 - v_{10}$ are the numbers of plane voxels that are divided into nine classes. u_{11} and v_{11} are the numbers of the other voxels.

From the feature descriptors $\boldsymbol{U}$ and $\boldsymbol{V}$, LPI is given by

$$\mathrm{LPI} = \frac{\sum_{i=1}^{11} \{\max(u_i, v_i) - |u_i - v_i|\}}{\sum_{i=1}^{11} \max(u_i, v_i)} \qquad (12)$$

A higher degree of similarity between the LiDAR scan data at both visit nodes leads to a larger LPI. Thus, the loop closure can be detected from the candidate of the revisit nodes using a large LPI value (a threshold of 80% in this paper). However, the LPI often fails in loop closure detection.



(a) Submap 1
(b) Submap 2
(c) Merged map

Figure 6. Submap generation and merging (top view).

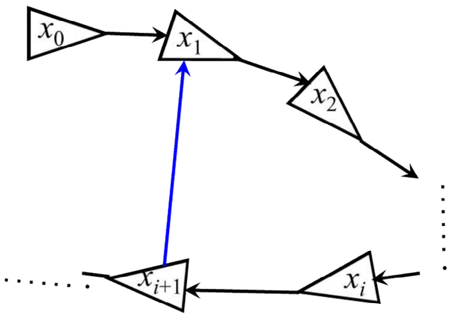


Figure 7. Pose graph in submap generation.

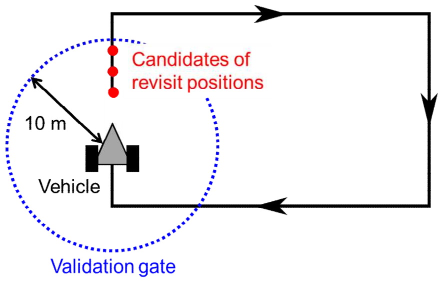


Figure 8. Loop closure detection in submap generation.

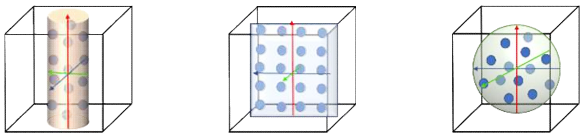

(a) Line voxel
(b) Plane voxel
(c) Other voxel

Figure 9. Classification of voxel.

The detection performance is then improved using a Matching Distance Indicator (MDI). From two LiDAR scan data captured at the current node and each candidate of the revisit nodes, the relative vehicle's pose is calculated based on NDT scan matching; the displacement of the self-locations at two nodes obtained by NDT SLAM is used as the initial relative pose for NDT scan matching.

In our experience, even if the relative pose of the vehicles at two nodes is large, a larger voxel size leads to a more robust matching in NDT scan matching. Therefore, the relative pose is calculated using two different voxel sizes. The relative pose is first calculated using a voxel size of 3 m. The obtained relative pose is used as the initial pose to calculate the relative pose by NDT scan matching with a voxel size of 1 m. The final estimate of the relative pose is applied to calculate the nearest neighbor distance between the two LiDAR scan data via NDT scan matching. The MDI is then calculated as

$$\mathrm{MDI} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{13}$$

where N is the number of measurements in the LiDAR scan data captured at the candidate of the revisit nodes. d_i is the nearest neighbor distance.

A higher degree of similarity between the LiDAR scan data captured at two nodes leads to a smaller MDI. The loop closure can then be detected by a smaller MDI value (a threshold of 1.5 m in this paper).

When the loop closures are detected by both LPI and MDI, the current vehicle's pose relative to its pose at the revisit node is inputted to the pose graph as a loop closure constraint (blue arrow in Figure 7). The objective function of (14) is then minimized to improve the accuracy of submap generated by NDT SLAM:

$$\begin{aligned} J(\chi) = &\sum_{i} \{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) - \boldsymbol{\delta}_{i+1,i}\}^T \boldsymbol{\Omega}^{pose} \{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) - \boldsymbol{\delta}_{i+1,i}\} \\ &+ \sum_{x_A, x_B \in \mathrm{loop}} \{(\boldsymbol{x}_B - \boldsymbol{x}_A) - \boldsymbol{\delta}_{A,B}\}^T \boldsymbol{\Omega}^{loop} \{(\boldsymbol{x}_B - \boldsymbol{x}_A) - \boldsymbol{\delta}_{A,B}\} \end{aligned} \tag{14}$$

where the first and second terms on the right side indicate the constraints on NDT SLAM and loop closure, respectively. $\boldsymbol{\chi} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_i^T, \cdots)^T$. $\boldsymbol{x}_i$ is the vehicle's pose at time $i\tau$. $\boldsymbol{\delta}_{i+1,i}$ is the relative pose of the vehicle between $i\tau$ and $(i+1)\tau$, which is calculated from NDT SLAM. $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are the vehicle's poses at the revisit and current nodes, respectively. $\boldsymbol{\delta}_{A,B}$ indicates the relative pose of the vehicle at the two nodes, which is calculated from the LiDAR scan data using NDT scan matching. $\boldsymbol{\Omega}^{pose}$ and $\boldsymbol{\Omega}^{loop}$ are the information matrices; they are inverse covariance matrices of NDT SLAM and given based on [28].

In this paper, we apply the open-source software g2o [29] to generate pose graphs and optimize (14).

B. Submap Merging

We consider the merging of submaps 1 and 2 in Figure 6. Submap merging is performed by the following steps:

- Loop closure detection: detection of encounter nodes in pose graphs corresponding to the two submaps;
- Relative pose estimation: estimation of the relative pose of the two submaps using the LiDAR scan data at nodes encountered in the two pose graphs;
- Alignment: coordinate transform of submap 2 using the relative pose estimate to represent the submap in Σ_{W1}; and
- Merging: merging of the two submaps using pose graph optimization.

If there are three or more submaps, an enlarged submap is first made by merging the two submaps, and another submap is then merged with the enlarged submap. By repeating such process, three or more submaps can be merged.

The loop closures between submaps (intersession loop closures) are detected based on LPI and MDI. However, unlike the loop closure detection in each submap (intrasession loop closure), the self-location information of the vehicle estimated by NDT SLAM is not useless in narrowing down the candidate of the encounter nodes in the two pose graphs because two submaps are generated in different coordinate frames. It is thus assumed that all nodes in the two pose graphs are the candidate of the encounter nodes, and the LPI is calculated by the brute force method to narrow down the candidate of the encounter nodes. Therefore, if the numbers of nodes are N_1 and N_2 in the pose graphs corresponding to submaps 1 and 2, respectively, the LPI is calculated $N_1 \times N_2$ times.

As shown in Figure 10, we consider that two nodes (the i-th node in pose graph 1, which corresponds to submap 1, and the j-th node in pose graph 2, which corresponds to submap 2) are detected as the candidate of the encounter nodes by the LPI (a threshold of 80%). To determine using the MDI whether or not the candidate is that of the encounter nodes, the relative pose of the vehicle is calculated from two scan data in both nodes via NDT scan matching. However, since two submaps are generated in different local coordinate frames, the initial pose, which is used to accurately calculate the relative pose by NDT scan matching, is unknown.

At the j-th node in pose graph 2, the vehicle coordinate frame Σ_b, in which the LiDAR scan data are captured, is

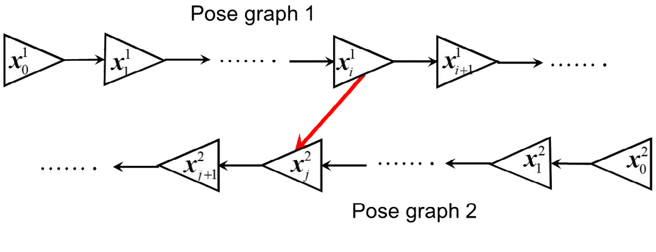


Figure 10. Pose graph in submap merging.

rotated about the z_b axis (yaw angle direction) in steps of 10° from 0° to 360° to address the above-mentioned problem. From the scan data at the *i*-th node in pose graph 1 and each of the 35 scan data at the *j*-th node in pose graph 2, NDT scan matching with an initial pose of zero value is applied to calculate the relative pose. If the relative pose estimate is correct, the MDI value will be small. The MDI for each of the 35 relative poses is then calculated, and the minimum MDI is selected. If the minimum MDI is 1.5 m or less, the candidate of the encounter nodes, the *i*-th and *j*-th nodes, is recognized as encounter nodes, and the relative pose is determined.

Such detection of intersession loop closure and related relative-pose calculation are repeated for all nodes in pose graphs 1 and 2. When many encounter nodes are detected in their pose graphs, the relative pose of the two pose graphs is determined by the weighted average of many relative poses. Using the relative pose, the coordinate transform of submap 2 is performed; consequently submaps 1 and 2 could be represented in the coordinate frame Σ_{W1}.

Finally, the relative pose of the vehicle at the encounter nodes is inputted to the pose graphs as the loop closure constraint (red arrow in Figure 10). The following objective function is then minimized to merge the two submaps:

$$J(\boldsymbol{\chi}^{total}) = J(\boldsymbol{\chi}^{1}) + J(\boldsymbol{\chi}^{2}) + \sum_{x_1, x_2 \in \mathrm{loop}} \{(\boldsymbol{x}_2 - \boldsymbol{x}_1) - \boldsymbol{\delta}_{1,2}\}^T \boldsymbol{\Omega}_{1,2}^{loop} \{(\boldsymbol{x}_2 - \boldsymbol{x}_1) - \boldsymbol{\delta}_{1,2}\} \qquad (15)$$

where $\boldsymbol{\chi}^{total} = (\boldsymbol{\chi}^{1T}, \boldsymbol{\chi}^{2T})^T$. $\boldsymbol{\chi}^{1} = (\boldsymbol{x}_1^{1T}, \boldsymbol{x}_2^{1T}, \cdots, \boldsymbol{x}_i^{1T}, \cdots)^T$ and $\boldsymbol{\chi}^{2} = (\boldsymbol{x}_1^{2T}, \boldsymbol{x}_2^{2T}, \cdots, \boldsymbol{x}_i^{2T}, \cdots)^T$ are sets of the vehicle's poses in pose graphs 1 and 2, respectively. $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$ are the vehicle's poses at times $i\tau$ and $j\tau$, respectively. $J(\boldsymbol{\chi}^{1})$ and $J(\boldsymbol{\chi}^{2})$ are the objective functions of the pose graphs corresponding to submaps 1 and 2, respectively. The third term on the right side is the constraint on the vehicle's relative pose in the merging of the two pose graphs. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are encounter nodes in pose graphs 1 and 2, respectively. $\boldsymbol{\delta}_{1,2}$ indicates the relative pose of the vehicle at the encounter nodes, which is calculated from the LiDAR scan data captured at the nodes using NDT scan matching. $\boldsymbol{\Omega}_{1,2}^{loop}$ is the information matrix; it is inverse covariance matrix of NDT scan matching and given based on [28].

VII. EXPERIMENTAL RESULTS

The performance of two methods is first examined, namely, distortion correction of LiDAR scan data and extraction of static scan data from the entire LiDAR scan data, which are presented in Sections IV and V, respectively. Thereafter, the mapping performance is shown through experimental results in residential and urban environments.

A. Performance of Distortion Correction of LiDAR Scan Data and Extraction of Static Scan Data

The experimental vehicle moves at a speed of about 40 km/h in two areas, as shown in Figures 11 (a) and (b). For comparison, the LiDAR scan data are mapped using NDT SLAM in the following cases:

Case 1: Mapping through the distortion correction of the LiDAR scan data and extraction of the static scan data from the entire scan data;

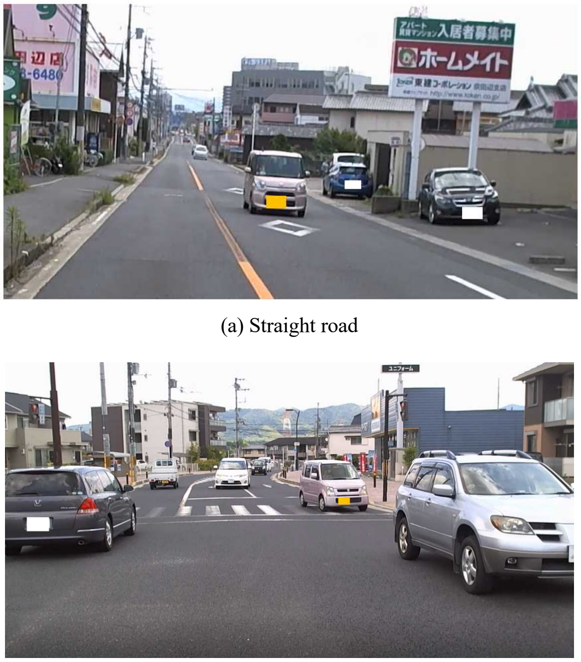
Case 2: Mapping without using either method.

Figures 12 and 13 show the mapping results on a straight road and an intersection area, respectively. The red line in (a) indicates the movement path of the experimental vehicle. The black and red dots in (b) and (c) indicate the static and moving scan data, respectively. These figures indicate that the extraction method of static scan data more significantly removes the tracks of cars. In the intersection, several cars slow down and stop at a red light or pause when turning left; they are determined as static objects. Consequently, in Figure 13 (b), LiDAR scan data related to cars partially remain.

Figure 14 shows the mapping result of a traffic sign in the road environment shown in Figure 12 (a). Figures 12–14 show that the mapping result obtained using the distortion correction of the LiDAR scan data is crisper than that obtained without using the distortion correction.

B. Mapping Performance

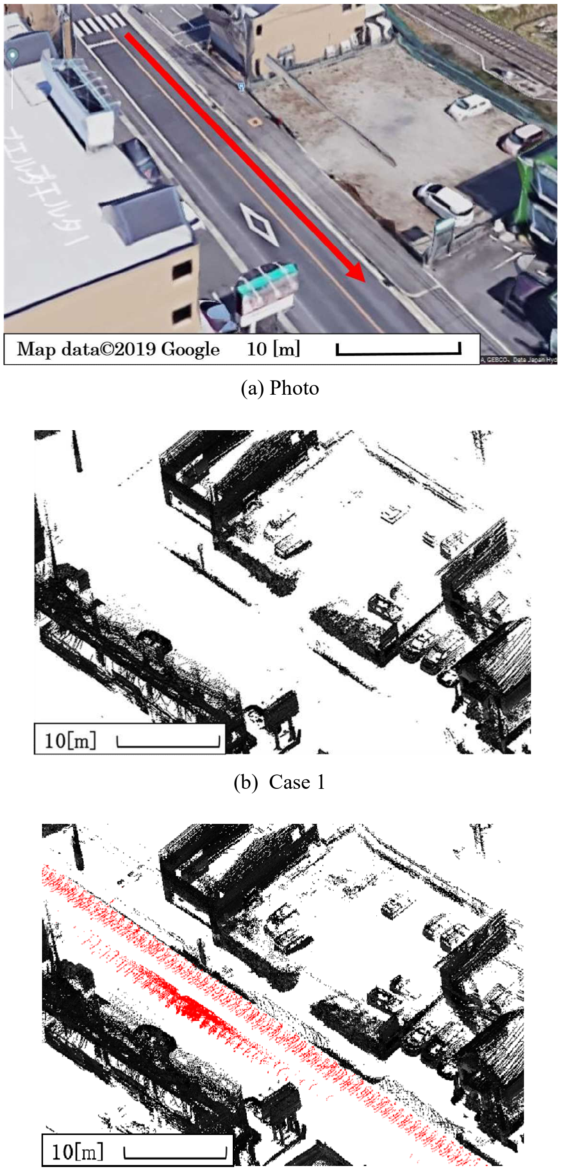
A mapping experiment is conducted in a residential environment near our university campus. The experimental vehicle moves at a speed of 10–20 km/h on a narrow road (6 m width) in the residential environment shown in Figure 15, and sensor data are recorded. The traveled distance of the vehicle is 2000 m. In Figure 15, the red point indicates the



(a) Straight road

(b) Intersection

Figure 11. Photo of experimental environment.



(a) Photo

(b) Case 1

(c) Case 2

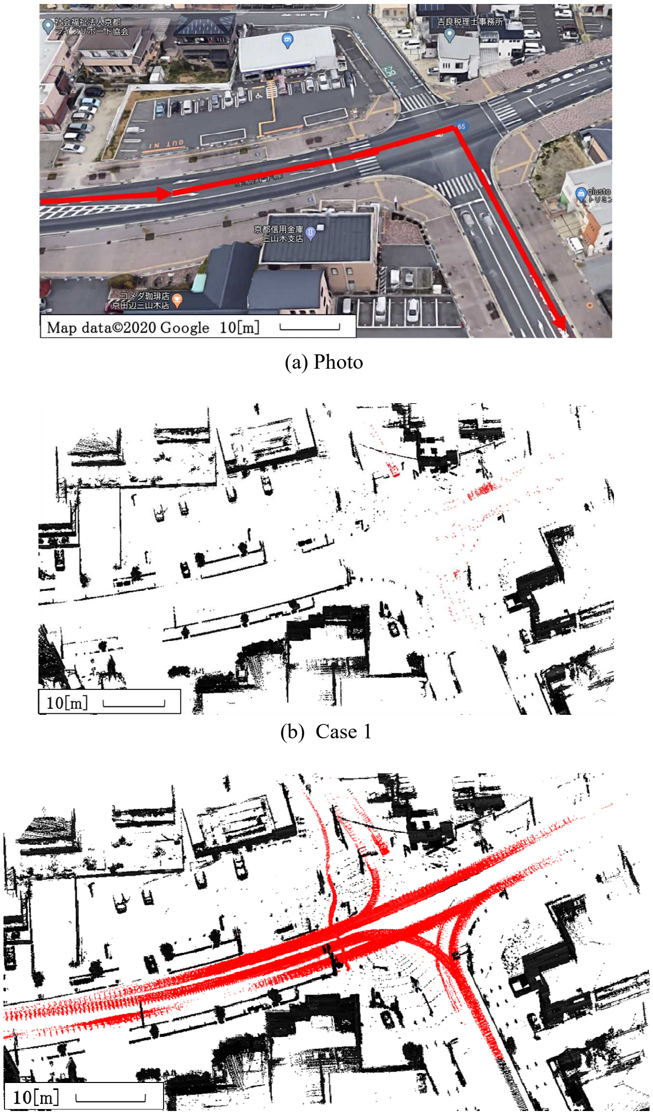
Figure 12. Mapping result of straight road area (bird's-eye view).

start/goal position of the vehicle. The black, blue, and green lines indicate the movement paths of the vehicle in areas 1, 2, and 3, respectively. The broken-line circles indicate the locations, at which areas 1, 2, and 3 overlap.

Figure 16 shows photos of the start/goal position and intersections 1 and 2, which are shown in Figure 15. In the residential environment, there are three cars and three pedestrians. One of the three cars always follows the experimental vehicle.

For comparison, maps are generated in the following cases:

Case 1: NDT-SLAM-based single-session mapping (single map generation) through the distortion correction of

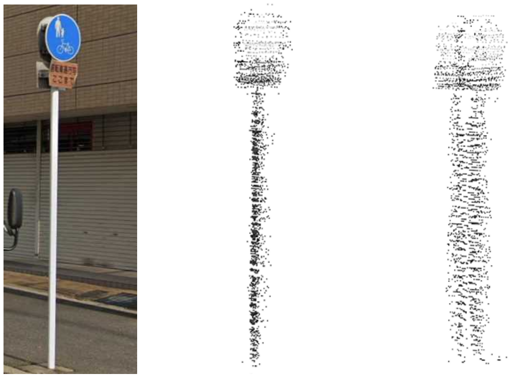


(a) Photo

(b) Case 1

(c) Case 2

Figure 13. Mapping result of intersection area (bird's-eye view).



(a) Photo

(b) Case 1

(c) Case 2

Figure 14. Mapping result of traffic sign.

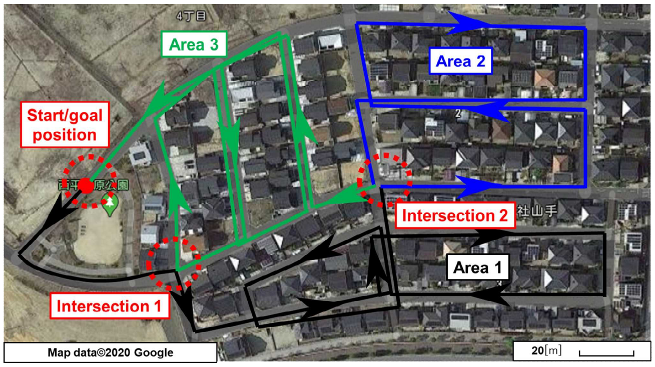
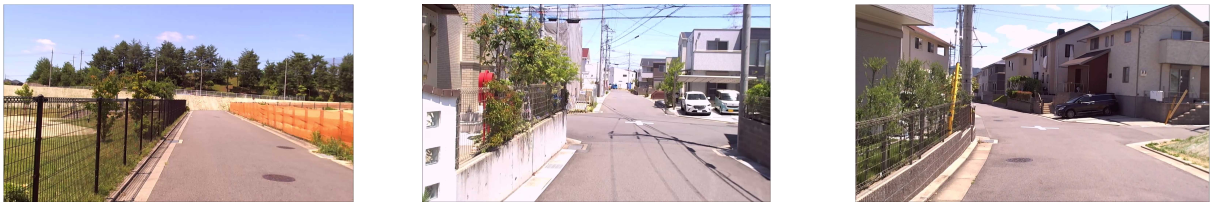


Figure 15. Movement path of vehicle (top view).

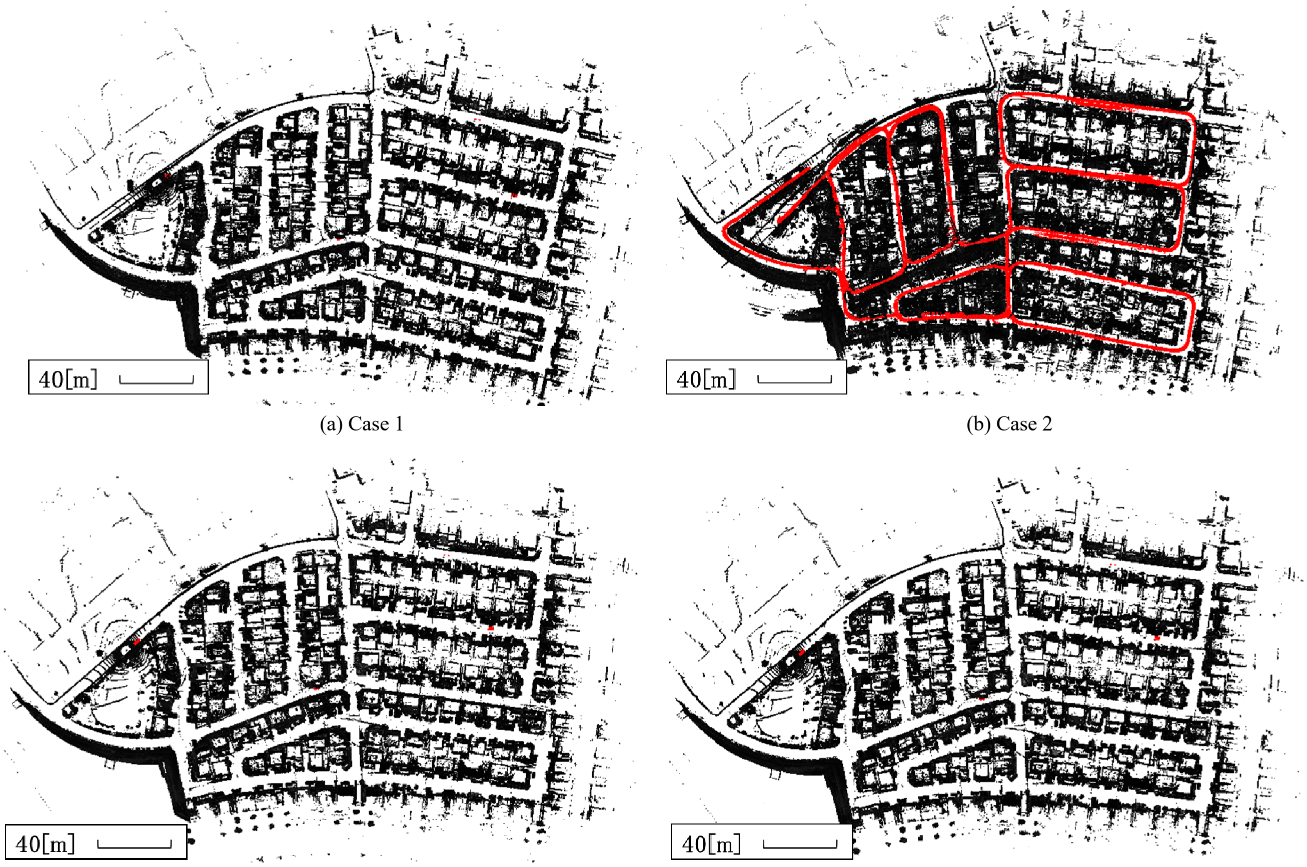


(a) Start/goal position

(b) Intersection 1

(c) Intersection 2

Figure 16. Photo of residential environment.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 17. Mapping results (top view).

the LiDAR scan data and extraction of the static scan data from the LiDAR scan data;

Case 2: NDT-SLAM-based single-session mapping without using either method;

Case 3: NDT-Graph-SLAM-based single-session mapping using both methods;

Case 4: NDT-Graph-SLAM-based multisession mapping (submap generation and mapping) using both methods (proposed method).

For case 4, we split the recorded LiDAR scan data into three segments that are assumed to be created independently in the three areas (areas 1, 2, and 3) shown in Figure 15. We then generate and merge three submaps using the split LiDAR scan data. The experimental vehicle moves approximately 700 m, 600 m, and 700 m in areas 1, 2, and 3, respectively. These three areas overlap at the start/goal position and intersections 1 and 2 in Figure 15. Submaps 1 and 2 are firstly merged, and their enlarged submaps are further merged with submap 3.

Figure 17 shows the mapping results in cases 1–4, where the black and red dots indicate the static and moving scan data, respectively. In case 3, 2799 revisit nodes are detected, and the map generated by NDT SLAM is modified. In case 4, the numbers of detected encounter nodes are 284, 543, and 1486 for submaps 1, 2, and 3, respectively. 24 encounter nodes are detected when submap 1 is merged with submap 2. Then, 1287 encounter nodes are detected when the enlarged submaps are further merged with submap 3.

As seen in Figure 17, although the mapping performance in case 2 is the worst, the difference in the mapping performance in cases 1, 3, and 4 is unclear due to the small scale of the map. In SLAM, the worse the performance of the self-location of the vehicle, the worse the mapping performance. Therefore, to quantitatively evaluate the mapping performance, we obtain the estimate error in the vehicle self-location estimated by SLAM, where position information from the onboard GNSS/INS unit is used as the ground truth of the vehicle.

Table I shows the deviation between the start and goal positions of the vehicle. Table II also shows the Root-Mean-Square Error (RMSE) of the self-location in the entire movement path of the vehicle. It is concluded from these tables that case 3 (single-session NDT-Graph SLAM) and case 4 (multisession NDT-Graph SLAM) provide better results than cases 1 and 2 (single-session NDT SLAM) do.

In the experiment in the residential environment, moving objects, such as cars and pedestrians, are very few. An experiment in an urban road environment is further conducted to show the mapping performance of the proposed method in dynamic environments.

The movement path of the vehicle and photo of the environment are shown in Figures 18 and 19, respectively. The traveled distance of the experimental vehicle is about 2900 m, and the maximum speed of the vehicle is 40 km/h. In the road environment, there are 114 cars, 26 two-wheelers, and 37 pedestrians.

For comparison, maps are generated in the four above-mentioned cases. For case 4, we split the recorded sensor data into three segments that are assumed to be created independently in the three areas (areas 1, 2, and 3) shown in Figure 18. We then generate and merge three submaps using the split sensor data. The vehicle moves approximately 900 m, 1100 m, and 900 m in areas 1, 2, and 3, respectively. These three areas overlap at intersection 1 in Figure 18.

Submaps 1 and 2 are firstly merged, and their enlarged submaps are further merged with submap 3. In case 3, 306 revisit nodes are detected, and the map generated by NDT SLAM is modified. In case 4, the numbers of detected encounter nodes are zero, 39, and zero for submaps 1, 2, and 3, respectively, because areas 1 and 3 are straight roads. 24 encounter nodes are detected when submap 1 is merged with submap 2. Then, 977 encounter nodes are detected when the enlarged submaps are further merged with submap 3.

Figure 20 shows the mapping result, where the black and green dots indicate the static scan data extracted in areas 1 and 3, respectively. The red dot indicates the moving scan data. Tables III and IV show the self-location results of the vehicle, which are estimated by SLAM.

As seen in Figure 20, the tracks of cars remain in case 2 because we do not implement the algorithm that removes the moving scan data from the entire LiDAR scan data.

TABLE I. DEVIATION BETWEEN START AND GOAL POSITIONS OF VEHICLE IN RESIDENTIAL ENVIRONMENT.

TRUE	*CASE 1*	*CASE 2*	*CASE 3*	*CASE 4*
12.31 m	14.43 m	32.40 m	12.12 m	11.75 m

TABLE II. RMSE OF SELF-LOCATION OF VEHICLE IN RESIDENTIAL ENVIRONMENT.

CASE 1	*CASE 2*	*CASE 3*	*CASE 4*
1.48 m	9.86 m	1.00 m	0.99 m

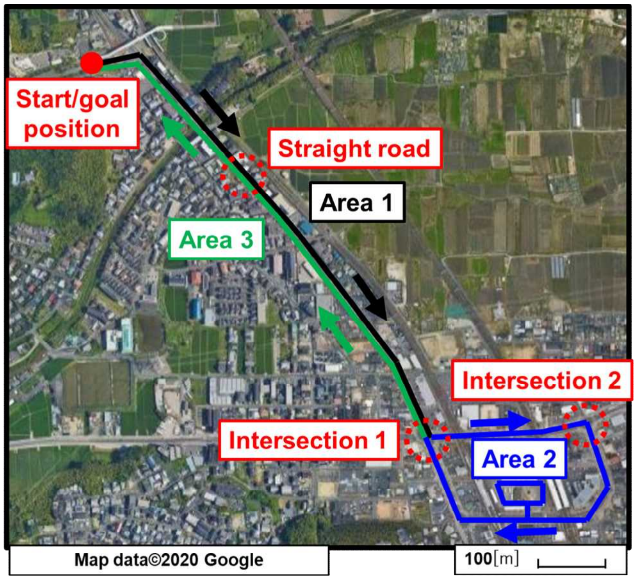
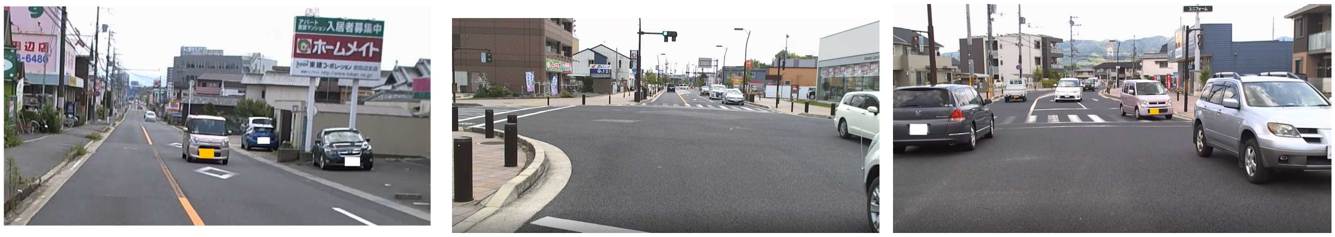


Figure 18. Moved path of vehicle (top view).

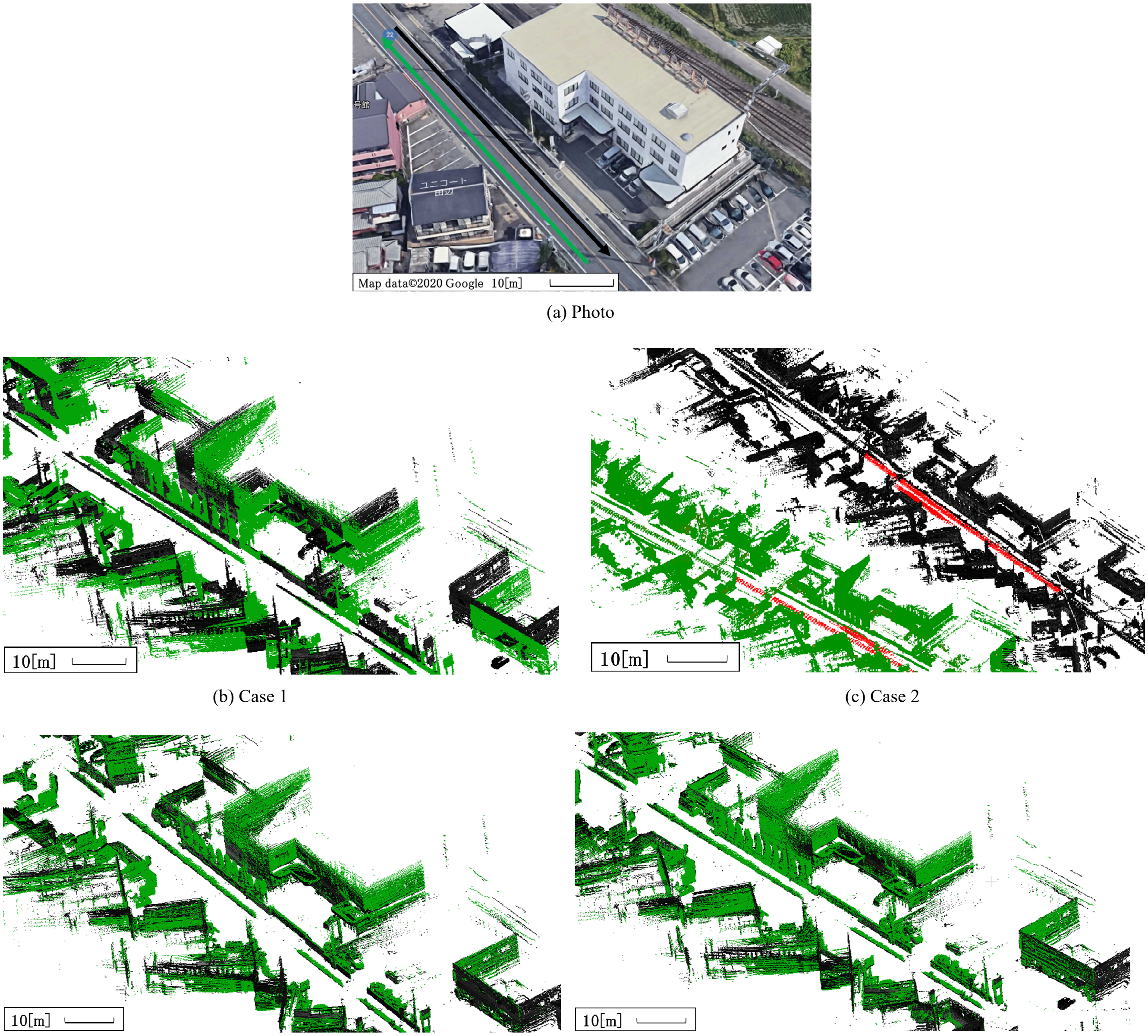


(a) Straight road

(b) Intersection 1

(c) Intersection 2

Figure 19. Photo of urban road environment.



(a) Photo

(b) Case 1

(c) Case 2

(d) Case 3

(e) Case 4

Figure 20. Mapping results (bird's-eye view).

TABLE III. Deviation between Start and Goal Positions of Vehicle in Urban Road Environment

True	*Case 1*	*Case 2*	*Case 3*	*Case 4*
3.50 m	17.34 m	126.19 m	6.10 m	3.38 m

TABLE IV. RMSE of Self-location of Vehicle in Urban Road Environment

Case 1	*Case 2*	*Case 3*	*Case 4*
5.95 m	35.95 m	9.86 m	3.23 m

Case 2 also causes a large drift in mapping due to the distortion of the LiDAR scan data and the accumulation error of NDT SLAM. The drift in case 1 is smaller than that in case 2 because the distortion of the LiDAR scan data is corrected in case 2. When the traveled distance of the vehicle is long, the accumulation error of NDT SLAM often deteriorates the performance of loop closure detection in Graph SLAM. For this reason, as seen in Table IV, the self-location error in case 3 (single-session NDT-Graph SLAM) is worse than that in case 1 (single-session NDT SLAM). Case 4 (proposed method) provides the best performance because shortly traveled distances in submaps reduce the accumulation error of NDT SLAM.

VIII. Conclusion and Future Work

This paper presented a method of LiDAR-based mapping and merging in GNSS-denied and dynamic environments using only an onboard scanning LiDAR. 3D point cloud mapping and merging were performed by integrating three previously proposed algorithms: distortion correction of LiDAR scan data, extraction of static scan data (removal of moving scan data) from the entire LiDAR scan data, and single-session and multisession mapping using NDT-Graph SLAM. The mapping performance was shown through experiments conducted in outdoor residential and urban road environments.

We are currently evaluating the proposed method by mapping various environments, including large-scale residential environments. Some improvements to the presented method are required. Since the distortion correction of the LiDAR scan data requires a great deal of computational time, Graphical Processing Unit (GPU) or Field-Programmable Gate Array (FPGA) must be utilized in real-time operations. In our method of moving-object detection, when, for example, cars slow down at an intersection, stop at a red light, or pause to turn left (or right), they are sometimes determined as static objects. Then, the LiDAR scan data that relate to cars partially remain on the environmental map. To address this problem, study on map update and maintenance is needed.

Acknowledgments

This study was partially supported by the KAKENHI Grant #18K04062, the Japan Society for the Promotion of Science (JSPS).

References

[1] M. Yamaji, S. Tanaka, M. Hashimoto, and K. Takahashi, “Point Cloud Mapping Using Only Onboard Lidar in GNSS Denied and Dynamic Environments,” Proc. of the 15th Int. Conf. on Systems (ICONS 2020), pp. 43–49, 2020.

[2] C. Cadena et al., “Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age,” IEEE Trans. on Robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[3] L. Wang, Y. Zhang, and J. Wang, “Map-Based Localization Method for Autonomous Vehicles Using 3D-LIDAR,” IFAC-Papers OnLine, vol. 50, issue 1, pp. 276-281, 2017.

[4] B. Huang, J. Zhao, and J. Liu. “A Survev of Simultaneous Localization and Mapping,” eprint arXiv:1909.05214, 2019.

[5] S. Kuutti et al., “A Survey of the State-of-the-Art Localization Techniques and Their Potentials for Autonomous Vehicle Applications,” IEEE Internet of Things Journal, vol.5, pp. 829–846, 2018.

[6] H. G. Seif and X. Hu, “Autonomous Driving in the iCity—HD Maps as a Key Challenge of the Automotive Industry,” Engineering, vol. 2, pp. 159–162, 2016.

[7] K. Morita, M. Hashimoto, and K. Takahashi, “Point-Cloud Mapping and Merging Using Mobile Laser Scanner,” Proc. of the third IEEE Int. Conf. on Robotic Computing (IRC 2019), pp. 417–418, 2019.

[8] D. Schwesinger, A. Shariati, C. Montella, and J. Spletzer, “A Smart Wheelchair Ecosystem for Autonomous Navigation in Urban Environments,” Autonomous Robot, vol. 41, pp. 519–538, 2017.

[9] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, “Simultaneous Localization and Mapping: A Survey of Current Trends in Autonomous Driving,” IEEE Trans. on Intelligent Vehicles, vol. 2, pp. 194–220, 2017.

[10] K. Inui, M. Morikawa, M. Hashimoto, and K. Takahashi, “Distortion Correction of Laser Scan Data from In-vehicle Laser Scanner Based on Kalman Filter and NDT Scan Matching,” Proc. of the 14th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO), pp. 329–334, 2017.

[11] S. Sato, M. Hashimoto, M. Takita, K. Takagi, and T. Ogawa, “Multilayer Lidar-Based Pedestrian Tracking in Urban Environments,” Proc. of IEEE Intelligent Vehicles Symp. (IV2010), pp. 849–854, 2010.

[12] S. Kanaki et al., “Cooperative Moving-Object Tracking with Multiple Mobile Sensor Nodes -Size and Posture Estimation of Moving Objects Using In-Vehicle Multilayer Laser Scanner-,” Proc. of 2016 IEEE Int. Conf. on Industrial Technology (ICIT 2016), pp. 59–64, 2016.

[13] S. Hong, H. Ko, and J. Kim, “VICP: Velocity Updating Iterative Closest Point Algorithm,” Proc. of 2010 IEEE Int. Conf. on Robotics and Automation (ICRA 2010), pp. 1893–1898, 2010.

[14] F. Moosmann and C. Stiller, “Velodyne SLAM,” Proc. of IEEE Intelligent Vehicles Symp. (IV2011), pp. 393–398, 2011.

[15] J. Zhang and A. Singh, “LOAM: Lidar Odometry and Mapping in Real-time,” Proc. of Robotics: Science and Systems, 2014.

[16] P. Biber and W. Strasser, “The Normal Distributions Transform: A New Approach to Laser Scan Matching,” Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2003), pp. 2743–2748, 2003.

[17] P. J. Besl and N. D. McKay, “A Method of Registration of 3-D Shapes,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239–256, 1992.

[18] W. Wen, L. T. Hsu, and G. Zhang, “Performance Analysis of NDT-Based Graph SLAM for Autonomous Vehicle in Diverse Typical Driving Scenarios of Hong Kong,” Sensors, 2018.

[19] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, “A Tutorial on Graph-Based SLAM,” IEEE Intelligent Transportation Systems Magazine, pp. 31–43, 2010.

[20] M. Labbe and F. Michaud, “Online Global Loop Closure Detection for Large-Scale Multi-Session Graph-Based SLAM,” Proc. of the 2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2014), 2014.

[21] J. McDonald et al., “6-DOF Multi-Session Visual SLAM Using Anchor Nodes,” Proc. of Int. Conf. on European Conf. on Mobile Robots (ECMR), pp. 69–76, 2011

[22] J. P. Saarinen, H. Andreasson, T. Stoyanov, and A. J. Lilienthal, “3D Normal Distributions Transform Occupancy Maps: An Efficient Representation for Mapping in Dynamic Environments,” Int. J. of Robotics Research, vol.32, no.14, pp. 1627–1644, 2013.

[23] X. Ding, Y. Wang, H. Yin, L. Tang, and R. Xiong, “Multi-Session Map Construction in Outdoor Dynamic Environment,” Proc. of the 2018 IEEE Int. Conf. on Real-time Computing and Robotics (IRC 2018), pp. 384–389, 2018.

[24] S. Thrun, “Learning Occupancy Grid Maps with Forward Sensor Models,” Autonomous Robots, vol.15, pp. 111–127, 2003.

[25] M. Munaro, F. Basso, and E. Menegatti, “Tracking People within Groups with RGB-D Data,” Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2012), pp. 2101–2107, 2012.

[26] R. B. Rusu and S. Cousins, “3D is Here: Point Cloud Library (PCL),” Proc. of the 2011 IEEE Int. Conf. on Robotics and Automation (ICRA 2011), 2011.

[27] F. Martín, R. Triebel, L. Moreno, and R. Siegwart, “Two Different Tools for Three-Dimensional Mapping: DE-based Scan Matching and Feature-Based Loop Detection,” Robotica, vol. 32, pp. 19–41, 2017.

[28] O. Bengtsson and A. J. Baerveldt, “Robot Localization Based on Scan-Matching—Estimating the Covariance Matrix for the IDC Algorithm,” Robotics and Autonomous Systems, vol. 44, pp. 29–40, 2003.

[29] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, “G2o: A General Framework for Graph Optimization,” Proc. of 2011 IEEE Int. Conf. on Robotics and Automation, pp. 3607–3613, 2011.

Active Monitoring Concepts for Safety-Critical Mirror Drivers of MEMS Micro-Scanning LiDAR Systems

Philipp Stelzer, Andreas Strasser, Philip Pannagger, Christian Steger
Graz University of Technology
Graz, Austria
Email: {stelzer, strasser, steger}@tugraz.at
pannagger@student.tugraz.at

Norbert Druml
Infineon Technologies Austria AG
Graz, Austria
Email: norbert.druml@infineon.com

***Abstract*—In the future, more and more cars will be equipped with Advanced Driver-Assistance Systems (ADAS) like Adaptive Cruise Control (ACC), Collision Avoidance System and many more. Currently, the driver is held responsible by law to perceive the environment and take over control if it is required. But in foreseeable future highly automated vehicles or even fully automated vehicles will appear on the road; where the vehicle is responsible for perceiving the environment, operating the vehicle and intervening in hazardous situations. By then it will be necessary that systems must not fail unnoticed. Therefore, it is mandatory to monitor safety relevant components. For instance Light Detection and Ranging (LiDAR) Systems like the 1D Micro-Electro-Mechanical System (MEMS) Micro-Scanning LiDAR, which will be part of intelligent sensor fusion in future ADAS. As a matter of course various safety monitors and safety devices are installed in highly automated vehicles to ensure an appropriately high level of safety. To further increase the safety level of the entire environmental perception system, we propose our novel Monitors for the Safety-Critical MEMS Driver of the LiDAR part in the sensor fusion unit. In this publication, we introduce novel system architectures that are able to verify the correct operation of internal control systems in MEMS-based LiDAR systems respectively to assess the reliability of the MEMS-based LiDAR in the sensor fusion unit of the entire environment perception system. To evaluate the effectiveness of our novel monitoring approaches, we implemented the procedures on a 1D MEMS Micro-Scanning LiDAR prototype platform.**

Keywords*–*ADAS; LiDAR; Signal Monitor; 1D MEMS Mirror; Safety Monitor

I. Introduction

With fully automated driving gaining more and more attention, industry and academia put a lot of effort into research in the field of sensor fusion and functional safety for sensors in the automotive domain. Key enablers of highly automated vehicles will be robust Radio Detection and Ranging (RADAR) and Light Detection and Ranging (LiDAR) solutions with additional support from vision cameras. Through fusion of sensor data and control functions enabling safe automated driving in rural as well as in urban environments is possible. In the project PRogrammable sYSTems for INtelligence in automobilEs (PRYSTINE) the consortium aims at a Fail-operational Urban Surround perceptION (FUSION) [2]. For years various Advanced Driver-Assistance Systems (ADAS), such as Electronic Stability Control (ESC) and Anti-lock Braking System (ABS) have been mandatory in new cars in the European Union [3]. ESC and ABS are ADAS, which are active safety components in contrast to passive safety components, such as seat belts and airbags [4]. For highly automated vehicles it is indispensable that ADAS are highly reliable and therefore ensure the safety for the driver, passengers and all other road users. Due to the increasing quantity and high reliability requirements of such ADAS and integrated systems the Society of Automotive Engineers (SAE) has introduced six levels of driving automation. A higher SAE level describes a higher level of driving automation of the vehicle. Due to the responsibilities that the systems take over the vehicle, it is possible to declare the SAE level of the vehicle [5]. Regardless of whether a vehicle, according to the manufacturer, would support higher automation levels, it is currently necessary in many countries that the driver continues to observe the environment and in an emergency takes over control [6]. For example, according to Article 8 of the Vienna Convention on Road Traffic, the driver must be able to continuously control the vehicle. The Vienna Convention on Road Traffic was ratified by the majority of EU member countries and several others. Large countries, such as the USA, China or England, are not among the signatories [7]. Due to legal and technical barriers driving automation levels

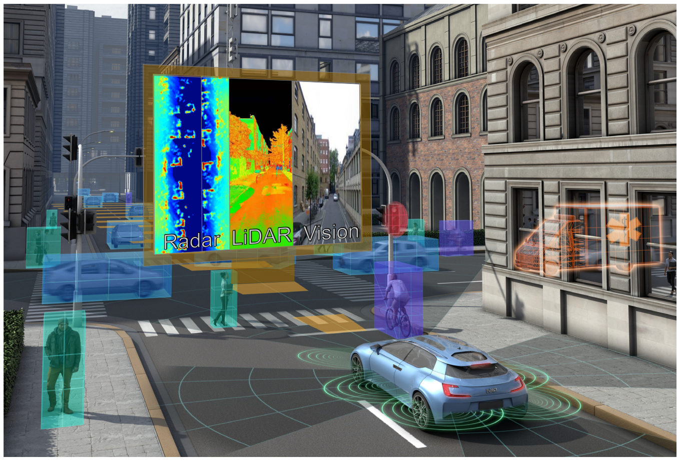


Figure 1. PRYSTINE concept view of a Fail-operational Urban Surround perceptION (FUSION) [2].

This publication is an extended Version of the “Monitor for Safety-Critical Mirror Drivers of MEMS Micro-Scanning LiDAR Systems” [1] publication, which was published in the Proceedings of the Tenth International Conference on Performance, Safety and Robustness in Complex Systems and Applications.

currently do not go beyond SAE level 2. From a legal point of view it will be necessary to adapt laws for introduction of vehicles with SAE Level 3 and greater in the future, as well as developing ADAS with higher levels of safety, reliability and availability. In projects like PRYSTINE, the goal is to develop components and systems for highly reliable and safe ADAS [2]. To ensure the proper functionality of systems it is mandatory to monitor said systems, especially parts which are safety critical. In case of a malfunction, the system has to be aware of its degraded state and in the worst case suspended its operation. Hence, these safety monitors are essential for ADAS in vehicles of SAE level 3 and above. Misbehaviour of a system is only detectable if the system is being monitored continuously. Therefore, we engaged in monitoring the Safety-Critical Mirror Driver of a 1D MEMS Micro-Scanning LiDAR System.

With our paper contribution we:

- Create a novel test opportunity for control loops.
- Ensure the detection of malfunctions during test run.
- Enable a reliability assessment of the LiDAR system.
- Allow for early warning about imminent failures of the LiDAR system.
- Enhance safety with diverse monitoring approaches.

Following aspects will be discussed: The overview on related work of MEMS-based LiDAR systems and several monitoring approaches are given in Section II. Architectures of novel safety monitors for the Safety-Critical Mirror Driver in a MEMS-based LiDAR System will be presented in detail in Section III and the achieved results including their discussion will be provided in Section IV. The summary and short discussion of the findings will conclude this paper in Section V.

II. Related Work

Currently available LiDAR technologies tend to be very bulky and cost intensive, such as the Velodyne HDL-64E [9]. Therefore, industry and academia put a lot of effort into the research of automotive qualified, long-range but low-cost LiDARs. Druml et al. introduced a 1D MEMS Micro-Scanning LiDAR, which is able to perceive the environment up to 200m, shall cost less than 200$ and is qualified for automotive applications due to its robustness [8]. The functional principle of the 1D MEMS-based LiDAR by Druml et al. is depicted in Figure 3. Several lasers are shot on the 1D MEMS mirror. A vertical laser beam is deflected by the mirror and sent into the scenery. This vertical line is moved horizontally across the Field-of-View(FoV) by oscillation of the mirror and the reflected light of the obstacle is captured by a stationary detector.

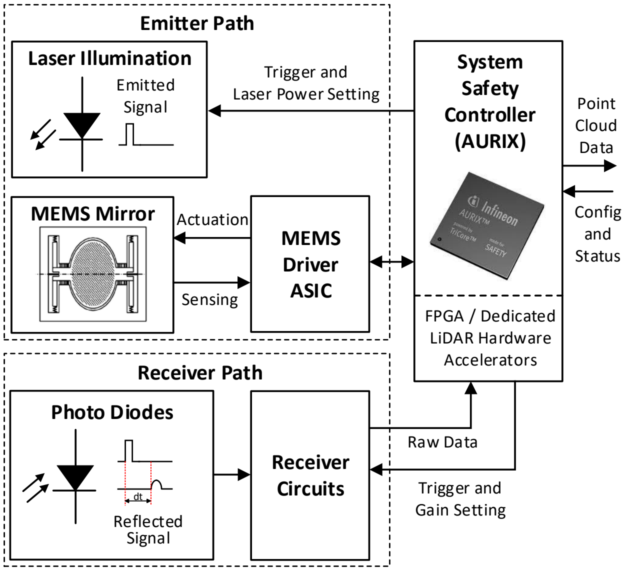


Figure 2. System concept of a 1D MEMS-based automotive LiDAR system by Druml et al. [8].

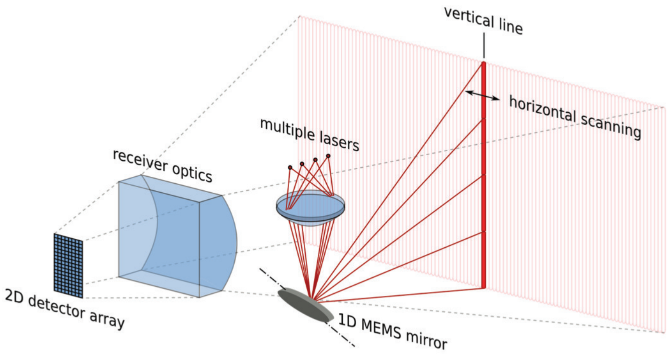


Figure 3. Functional principle of a 1D micro-scanning LiDAR [8].

A. 1D MEMS Micro-Scanning LiDAR

In this section, the 1D MEMS-based LiDAR System by Druml et al. is presented. The system concept of the MEMS-based LiDAR is depicted in Figure 2. Generally Druml et al.'s system consists of an emitter path, a receiver path and the System Safety Controller (AURIX). The emitter path includes a laser illumination unit, the MEMS mirror and the actuation and sensing unit of the mirror, the MEMS Driver ASIC. Within the receiver, an array of photo diodes and the receiver circuitry is included. The System Safety Controller is the central unit, which is responsible for monitoring, controlling and signal processing. Regarding the signal processing part, the task of the System Safety Controller is to compute and provide a 3D point cloud for dedicated ADAS [8]. Due to the dependence of correct position, direction and verification signals from the mirror, the Driver ASIC, which is responsible for the actuation and sensing of the MEMS mirror, is described in particular. The MEMS Driver provides crucial signals to the System Safety Controller. Thereby it is mandatory that the delivered information is reliable. By reference to the correctness of these crucial signals, the System Safety Controller will create a plausible 3D point cloud with the raw data from the receiver circuits. If the crucial signals were corrupted, the 3D point cloud would be useless due to wrong assumptions of the reflected laser origin.

In Figure 4, the crucial signals are illustrated, which are provided by the MEMS Driver ASIC. These signals are needed to monitor the current status of the MEMS mirror during operation. The POSITION_L represents whether the mirror is aligned to the left or to the right side; logical high means an alignment to the left and logical low to the right. DIRECTION_L indicates in which direction the movement is directed; logical high means moving to the left and logical low to the right. Precise and high-frequent phase information of the current mirror position is provided by a PHASE_CLK signal that counts from 0 to n_{max} in equal time steps during one mirror oscillation. Furthermore, an ANGLE_OK signal is available in addition to the tracking signals. This ANGLE_OK signal notifies the System Safety Controller when the Driver ASIC operates according to the programmed specification (e.g., angle setpoint is reached). To be able to ensure functional-, eye-, and skin-safety this notification is mandatory: MEMS mirror's current position and MEMS Driver ASIC's internal position information must match to allow the laser to be emitted [8].

B. Test Facilities

One of the major objectives of the automobile industry is to evolve individual traffic. The coexistence of partially, highly and fully automated cars will be the reality in the near future. In conventionally equipped vehicles, the driver is responsible for environmental perception, operation of the vehicle and intervention in hazardous situations. In prospective automated cars more and more competences will move from the driver to the car. Based on information, which is obtained from ADAS, the vehicle will make decisions. Therefore, it is obviously necessary that this information is reliable. To ensure safe and reliable operation of ADAS and their embedded components like LiDAR, it is mandatory to test the behavior for correctness. BISTs and a wide variety of safety monitors can be used for this purpose.

1) Built-In Self-Test:

A Built-In Self-Test (BIST) operates simultaneously with the circuit and is monitoring or checking the output of a circuit to check its validity. The BIST needs a strategy for generating input signals for the circuit and has to know how to evaluate the correlated output. The circuit or device which is tested is called the Circuit Under Test (CUT). A basic BIST architecture is shown in Figure 5. A realization of a BIST fundamentally needs to implement four new functions within the existing system. First of all, there is the Test Pattern Generator (TPG), which is responsible for generating the input signals for the test. The test pattern consists of multiple sets of test cases, which theoretically simulate all possible combinations of input signals. The complement to the TPG is the Output Response Analyzer (ORA). Its task is to know every correct output response of the CUT and decides whether the current output is faulty or valid. To create a meaningful and valid test it is important to isolate the test from any other input. Therefore, the Input Isolation Circuitry (IIC) is implemented. Its task is to decouple all input signals, which are commonly provided to the CUT and replace them with test-signal coming from the TPG. Last, but not least to synchronize the behaviour of the TPG, ORA and IIC the Test Controller is implemented. First it initializes a specific test, then decouples the System Inputs and finally activates the ORA which then outputs a Fail or Passed signal [10][11].

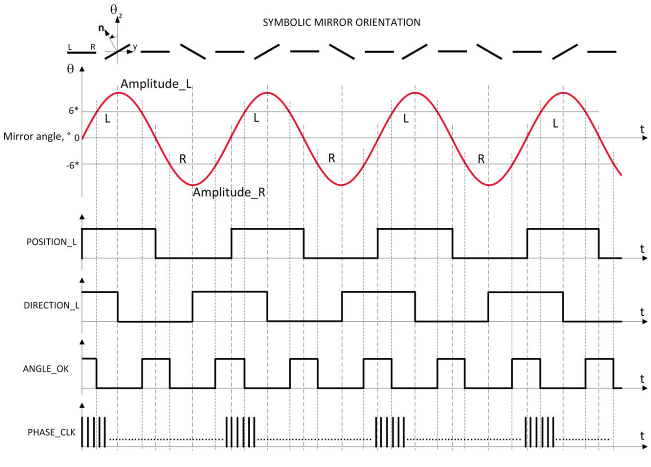


Figure 4. Crucial signals of the MEMS Driver ASIC from Druml et al.'s LiDAR system [8].

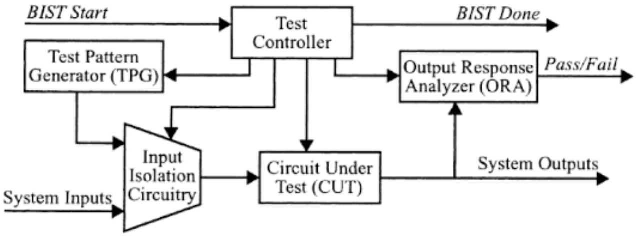


Figure 5. A basic Built-In Self-Test Architecture [10].

2) Safety Monitor Approaches:

Beside BISTs there are also other monitors, which verify the behavior of circuits and whole systems. Schuldt et al. [12], for example, strive to test and validate ADAS efficiently by referencing systematically generated virtual test scenarios. The idea hereby is to identify the factors that affect the assistance system. Hence, the test scenarios will be generated. By reference to the test scenarios a test will be executed and due to the variety of scenarios an evaluation of the results can be done. Another approach to monitor ADAS is presented by Mauritz et al. [13]. With this approach, results obtained from simulations are transferred to road scenarios. They ensure a consistent behavior of the ADAS in both worlds due to a simulation of realistic driving conditions and by utilization of a set of runtime monitors. Furthermore, Meany [14] illustrates that Integrated Circuits (IC) provide the basis for all modern safety-critical systems. According to Meany, besides redundant and diverse development, it is necessary to monitor the ICs to achieve fault-tolerance. There are several ways to monitor the IC during operation. Meany addresses several opportunities of IC diagnostics in his paper.

3) On-Board Diagnostic Systems:

The California Air Research Board (CARB) was established in 1967 as commission of experts to draw up legislative proposals for control of air pollution. The idea of On-Board Diagnostic (OBD) systems for vehicles was then born on the one hand by CARB and on the other hand by the car manufacturers themselves in the 1970s. McCord [15] explains in his book, how it came about from the establishment of this agency in 1967 to the OBD protocols that are standardised today. OBD-I, all standards before OBD-II was introduced, dealt with engine malfunctions and emission equipment malfunctions. OBD-I and OBD-II are well described in several publications [15], [16], [17]. In difference to the OBD-I regulations in which only a limited number of components had to be monitored, the current OBD-II regulations include monitoring of a wide range of components and systems that in turn also monitor components. The fundamental strategy behind the OBD II system is unchanged from the OBD-I system: OBD-II systems

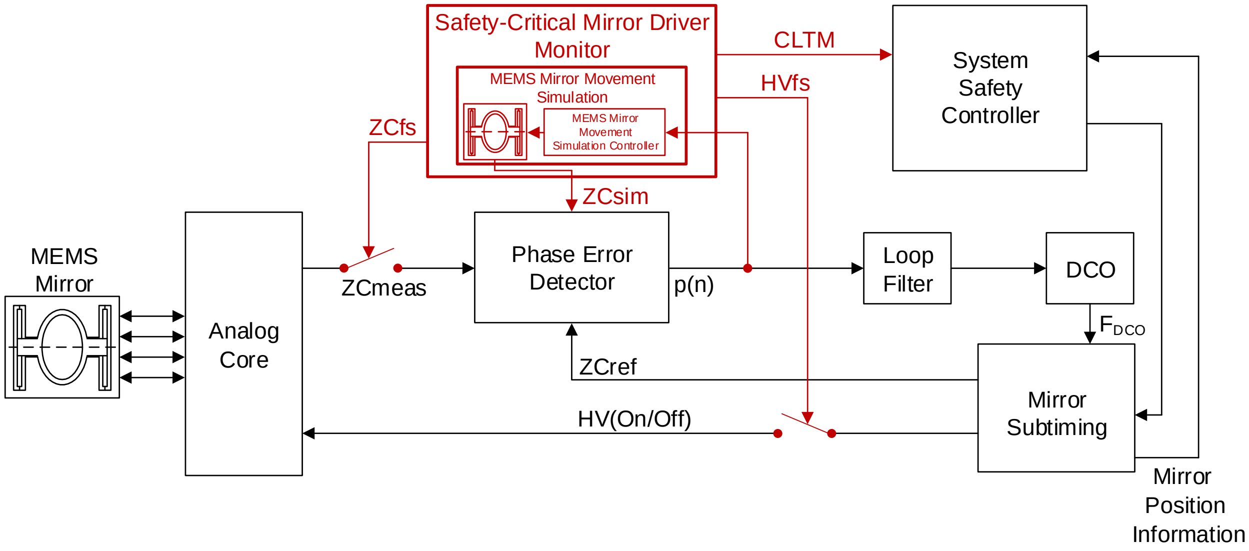


Figure 6. Block diagram of a PLL architecture with the novel adaptions to include a Safety-Critical Mirror Driver Monitor module in the system.

monitor emission-related components. When a problem is detected, the driver of the vehicle is alerted by the illumination of the so-called Malfunction Indicator Light (MIL) on the dashboard. The MIL can be triggered under a variety of conditions, as Durbin et al. [18] described in their publication. In the case of sporadic faults (e.g., due to a loose contact), the MIL may turn off after the fault has disappeared or after the next engine start. Otherwise, this can only be done by reading out and clearing the fault memory at the workshop. This approach can also be pursued for other monitors. For highly automated vehicles, a system degradation can be logged and further examined with similar approaches.

III. Core Concepts and Architectures

In this section, we present our concepts and architectures for novel safety monitors of MEMS-based LiDAR systems. The reliability of the Driver is a sensitive topic. Therefore, it is indispensable to monitor and test the Driver extensively and diverse. Thus, we introduced novel procedures to enable testing and monitoring the Driver component.

A. Novel Safety-Critical Mirror Driver Monitor

The first procedure, we present in our publication is a novel safety monitor for the Driver to check the functionality of the phase-locked loop (PLL) control. It is a procedure to evaluate the correct operation of a control loop, while the system is not in use. For deeper insight into this concept of the procedure, the architecture and process flow will be described in the following. At first, the architecture modifications are highlighted and described. Furthermore, we go through the process flow of the monitoring and test period. With this new monitor there is another possibility to detect faults in the Driver module at an early stage and to take appropriate measures beforehand. In case of detected faults, for example, the System Safety Controller will be informed and the LiDAR system can be degraded or disabled accordingly. Due to the diversity of the testing module it should be possible to prevent prior undetectable faults even better.

In Figure 6, the modified block diagram is illustrated. In principle, it is a common PLL, which is essential for the MEMS mirror actuation, the System Safety Controller, the MEMS mirror and our novel Safety-Critical Mirror Driver Monitor (SCMDM). The HV(On/Off) signal sets the points in time in the internal schedule at which the High Voltage (HV) is switched on or off. This internal schedule is managed by the Mirror Subtiming block. How fast or slow this schedule is processed depends on the PLL and therefore we aimed at testing the PLL on its functionality. For this purpose we designed a SCMDM and adapted the existing architecture and integrated our novel monitor. The core of the SCMDM consists of a mirror simulation part and a decision part. The decision part is responsible to evaluate the test run and notify the System Safety Controller. With the begin of the test run and the accompanying monitoring of the system, it is also necessary to decouple the Driver from the physical MEMS mirror. Hence switches for the Zero-Crossing measured (ZCmeas) and High Voltage On/Off (HV(On/Off)) signals were implemented. To start the test run the SCMDM block disables the switch for ZCmeas signal by Zero-Crossing forwarding stop (ZCfs) signal and the switch for HV(On/Off) signal by High Voltage forwarding stop (HVfs) signal. Furthermore, the SCMDM notifies the System Safety Controller of the test run by the Control Loop Test Mode (CLTM) signal.

After a test run is started the Zero-Crossing simulated (ZCsim) signal is forwarded to the Phase Error Detector (PD) block instead of the ZCmeas signal. A test run can be started at a vehicle startup or even while stopping in front of a traffic light. In case of a vehicle startup, the frequency of the simulated MEMS mirror movement is set to a random but plausible frequency. Otherwise, the frequency is set to a different frequency than the actual mirror swing to test and monitor the behaviour of the MEMS Driver during control operation. To be able to adapt the simulated frequency to the Zero-Crossing (ZC) a MEMS Mirror Movement Simulation Controller (MMMSC) is implemented in the simulation part of the SCMDM. By reference to the PLL error this controller is adapting the simulated MEMS mirror frequency and works contrary to the PLL. Due to the characteristics of the MEMS mirror in regard to acceleration and deceleration, the control loop of the simulation must take these into account. This is necessary to be able to emulate the physical MEMS mirror's behavior after frequency increase respectively decrease. The

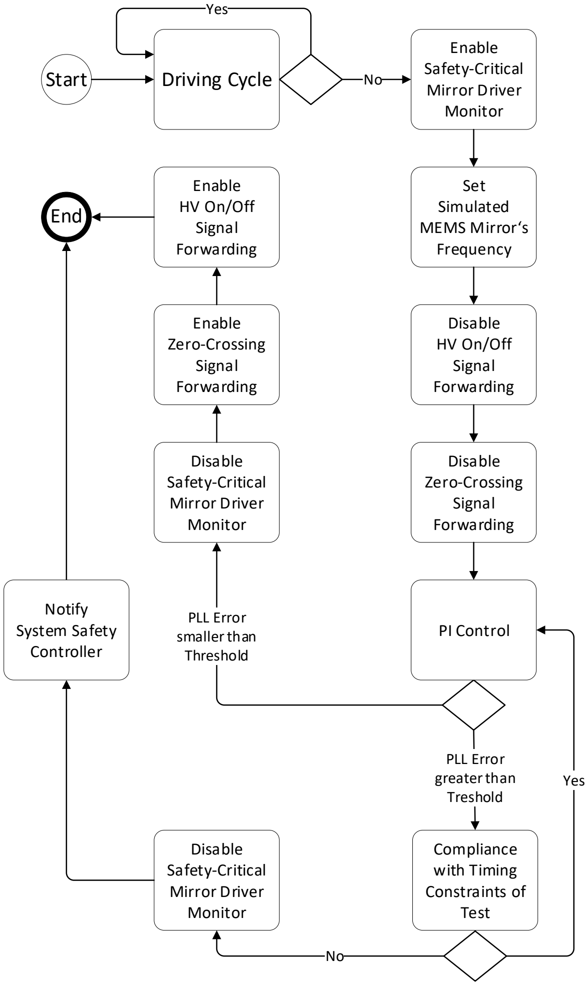

Yes
Start
Driving Cycle
No
Enable Safety-Critical Mirror Driver Monitor
Set Simulated MEMS Mirror's Frequency
Disable HV On/Off Signal Forwarding
Disable Zero-Crossing Signal Forwarding
PI Control
PLL Error greater than Treshold
Yes
Compliance with Timing Constraints of Test
No
Disable Safety-Critical Mirror Driver Monitor
Notify System Safety Controller
PLL Error smaller than Threshold
Disable Safety-Critical Mirror Driver Monitor
Enable Zero-Crossing Signal Forwarding
Enable HV On/Off Signal Forwarding
End

Figure 7. Process flow of the Safety-Critical Mirror Driver Monitor module.

acceleration of the mirror requires more energy effort than its deceleration. Thus, the integrator values have to be chosen accordingly to that fact. An overview of the process flow of this procedure is depicted in Figure 7. The test cycle and monitoring procedure is divided into the following steps:

1) **Checking for Driving Cycle**

The operational state of the vehicle is continuously examined whether the vehicle is in the driving or not. A stopped driving cycle is, for example, a vehicle stop before a traffic light or a vehicle start. A test cycle with subsequent mirror restart is usually shorter than one second. In both cases, traffic light stop and vehicle start, there is at least 1s time to perform the test and monitoring cycle. Hence, the SCMDM is started after a stop of the driving cycle is detected.

2) **Enable Safety-Critical Mirror Driver Monitor**
After the driving cycle check green lights the test the SCMDM is enabled and notifies the System Safety Controller via the CLTM signal about the test cycle. The next step is to adjust the frequency for the simulated mirror.

3) **Frequency Adjustment**
On the basis of a simulated mirror movement the adequate and orderly function of the MEMS Driver ASIC's PLL shall be proven. Therefore, it is necessary to set a start frequency for this simulated mirror with a significant difference to the actual frequency of the physical MEMS mirror. In case of a vehicle start it is only necessary to choose a frequency within given limits of the physical MEMS mirror. If the MEMS mirror has already been in operation, the frequency to be set must then be selected within plausible limits and the selected frequency must also be sufficiently different from the actual mirror frequency. After the initial frequency of the mirror simulation is set the system has to be decoupled from the physical MEMS mirror during the test cycle.

4) **Decoupling**
Switches have been integrated into the existing architecture to decouple the system from the MEMS mirror. By means of HVfs the HV(On/Off) signal is decoupled from the physical mirror and thus prevents an unintended mirror actuation. During the test phase, the mirror is actuated in an open loop mode with the HV(On/Off) value, which is configured before the test is started. In order to prevent a disturbance of the control loop during test mode by the ZC of the physical mirror, the ZCmeas signal is switched off. Thereby pnly the ZCsim signal is forwarded to the PD block and the PLL is not affected due to two different, actual and simulated ZC, signals.

5) **PI Control**
Afterwards the control of the PLL and the simulated mirror frequency starts. The PLL is operating as usual and tries to match the internal adjusted frequency with the simulated mirror frequency. The simulated mirror is also adapting the frequency with respect to the specifics of the acceleration and deceleration of the physical mirror. By reference to the obtained PLL error the MEMS Mirror Movement Simulation (MMMS) part is informed whether an acceleration (frequency increase) or a deceleration (frequency decrease) has to be simulated. It is necessary to know whether the simulated mirror needs to be accelerated or decelerated because the integrator values of acceleration and deceleration differ. Due to the difference in energy consumption between acceleration and deceleration. This regulation happens until either the simulated mirror has the desired frequency or a time limit is reached.

6) **End of PI Control**

a) **Control Success**
After the control process was successful, the SCMDM is disabled and the physical MEMS mirror is integrated into the control system again instead of the simulated one. To re-integrate the MEMS mirror, the ZCmeas signal is forwarded to the PD block and the HV(On/Off) signal of the Mirror Subtiming block is forwarded to the Analog Core that connects to the physical mirror.

b) **Control Abort**
In case the control is aborted by reaching the time limit, the SCMDM is also disabled. In contrast to successful control, however, a notification of failure is transmitted to the

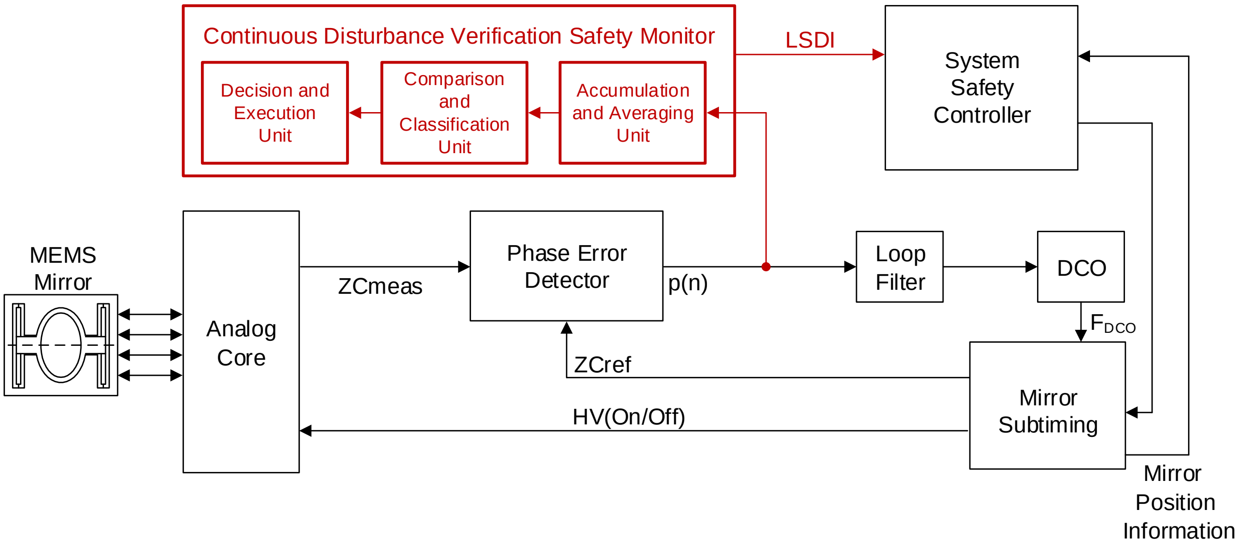

Continuous Disturbance Verification Safety Monitor
Decision and Execution Unit
Comparison and Classification Unit
Accumulation and Averaging Unit
LSDI
System Safety Controller
MEMS Mirror
Analog Core
ZCmeas
Phase Error Detector
p(n)
Loop Filter
DCO
F_DCO
ZCref
HV(On/Off)
Mirror Subtiming
Mirror Position Information

Figure 8. Block diagram of a PLL architecture with the novel adaptions to include a Continuous Disturbance Verification Safety Monitor module in the system.

System Safety Controller. The System Safety Controller is then responsible for further measures. Such measures could be a further test run or a degradation of the system.

7) **Encoupling**
After the test run is finished, the physical mirror is coupled back into the system. This works in principle similar to the start-up procedure. The physical mirror in open loop mode is put back into closed loop mode by activating the PLL. This completes the test run and the system continues to operate as usual.

With this novel procedure there is the possibility to check the function of a control loop for MEMS-based LiDAR systems. Especially for safety-critical components in environmental perception systems, it is important due provide diversity in addition to redundancy of tests and monitoring. The most important thing is to ensure the correct operation of the systems that provide information for ADAS and other sensor fusion components. Section IV discusses and explains the results of the novel monitor approach.

B. Continuous Disturbance Verification Safety Monitor

The second procedure, we present in our publication, is a novel safety monitor for the MEMS Driver to continuously check the system for disturbances. This procedure is focused on disturbances in the control loop, which can be detected via the provided PLL error during the system runtime. To obtain a more detailed understanding of this concept of the procedure, the architecture and process flow will be discussed in the following. First of all the architectural changes are highlighted and described. Furthermore, the process flow of the monitoring and degradation steps will be illustrated. Another possibility for disturbance detection and the corresponding degradation measures is made possible by this new type of monitor. For example, if a reoccurring disturbance is detected, measures can be taken depending on the severity of the disturbance, ranging from partial degradation to complete degradation of the LiDAR system. As a result it should be possible to degrade supposedly malfunctioning MEMS-based LiDAR systems in sensor fusion units of environment perception systems.

Figure 8 shows the block diagram of a common PLL architecture with the modifications for the integrated **Co**ntinuous **D**isturbanc**e** Ver**i**fication **S**afety M**o**nitor (CodeIso) and the System Safety Controller. The PLL is responsible for matching the frequencies of the MEMS mirror and the MEMS Driver. With a constant low PLL error, the frequencies of the MEMS mirror and MEMS Driver are approximately equal. If the PLL error increases, this may be due to several reasons. It can be caused, for example, by an frequency adaption during the adjustment phase to the new frequency or by a massive shock. Or due to physical problems with the MEMS mirror such as ageing or other signs of wear. Therefore, we designed a CodeIso and integrated this novel monitor into the existing architecture. The CodeIso is essentially composed of an Accumulation and Averaging Unit (AAU), a Comparison and Classification Unit (CCU) and a Decision and Execution Unit (DEU). The AAU is responsible for accumulating the absolute PLL error values over a specified number of Mirror Half Periods. These accumulated absolute PLL error values will afterwards be averaged and forwarded to the CCU. In the CCU, the PLL error mean value obtained will be compared with a PLL error mean value set by an authorised mechanic or technician during the last maintenance in the repair shop. Depending on the deviation of the obtained PLL error mean value from the preset PLL error mean value, the measurement is classified into a Degradation Level. The classified Degradation Level will then be stored as a histogram. This histogram is subsequently forwarded to the DEU to be able to validate the Overall Degradation Level of the LiDAR system. In the DEU a validation of the Overall Degradation Level takes place. According to the results of this validation, further action can be taken. In any case, the System Safety Controller will be informed of the Level of System Degradation Indicator (LSDI) of the current Degradation Level of the LiDAR system. The System Safety Controller is the interface between the LiDAR system and the sensor fusion unit in the entire environmental perception system. With this information the LiDAR system is then degraded by the System Safety Controller in the sensor fusion unit of the environment perception system when the LSDI indicates a necessary degradation. Such Degradation Levels can either change again during runtime or, under certain circumstances, only be altered after the system has been inspected, repaired if necessary respectively replaced and finally released. This monitor is used to observe the system and does not take corrective action. The purpose of this procedure is to ensure that any disturbances are detected and the environment perception system can be alerted accordingly. The procedural flow is depicted in Figure 9. The monitoring procedure is divided into the following steps:

1) **Checking for System Degradation**
After startup the LiDAR system first checks whether the system is fully degraded or not. If the system is fully degraded, the CodeIso is not enabled. The CodeIso can only be re-enabled after the system has been inspected, repaired respectively replaced and released. Otherwise, the CodeIso is started after the system startup and operates during the whole system runtime until the system is degraded or the system is shut down.
2) **Enable Continuous Disturbance Verification Safety Monitor**
When the degradation check shows that the system is not fully degraded and therefore not neglected in the sensor fusion, the CodeIso is enabled. The CodeIso is now active as long as there is no full system degradation. The system is monitored during operation by the CodeIso.
3) **PLL Error Accumulation**
As soon as the CodeIso is active, the absolute PLL error value accumulation starts. For each Mirror Half Period, a PLL error is measured that occurs between the actual ZC of the MEMS mirror and the ZC reference signal. This PLL error is then used as an absolute value to average the PLL error values over a certain measuring period and is cached. Until the desired number of PLL error values per Mirror Half Period is reached, these absolute PLL error values are constantly accumulated.
4) **Averaging Accumulated PLL Error**
After the accumulation of the absolute PLL error values is complete, the PLL Error Mean Value (PEMV) is formed.

$$PEMV = \frac{1}{n} \sum_{i=1}^{n} |PEV_i| \qquad (1)$$

In Equation (1), the PEMV is calculated by reference to the sum of the individual absolute PLL error values and the quantity of PLL error values. PEV$_i$ represents the PLL error value of measurement i. This PEMV is then forwarded to the CCU to compare and classify the state of the system.
5) **Compare and Classify PLL Error Mean Values**
During maintenance, the system is inspected and a mean value of the measured absolute PLL error values during proper operation is formed. This Maintenance PLL Error Mean Value (MPEMV) is compared with the previously calculated PEMV. Depending on the deviation from the MPEMV, the PEMV is classified into a Degradation Level.
6) **Creation of Histogram**
The histogram is afterwards filled with the previously classified Degradation Levels. Depending on the level of degradation, the entry in the histogram is weighted. For example, for Degradation Level 0, each Degradation Level 0 entry is increased by 1. For Degradation Level 1 it is increased by 1.5 and for Degradation Level 2 by 2. According to how significant a Degradation Level should be, you can change the weighting. The histogram is filled up

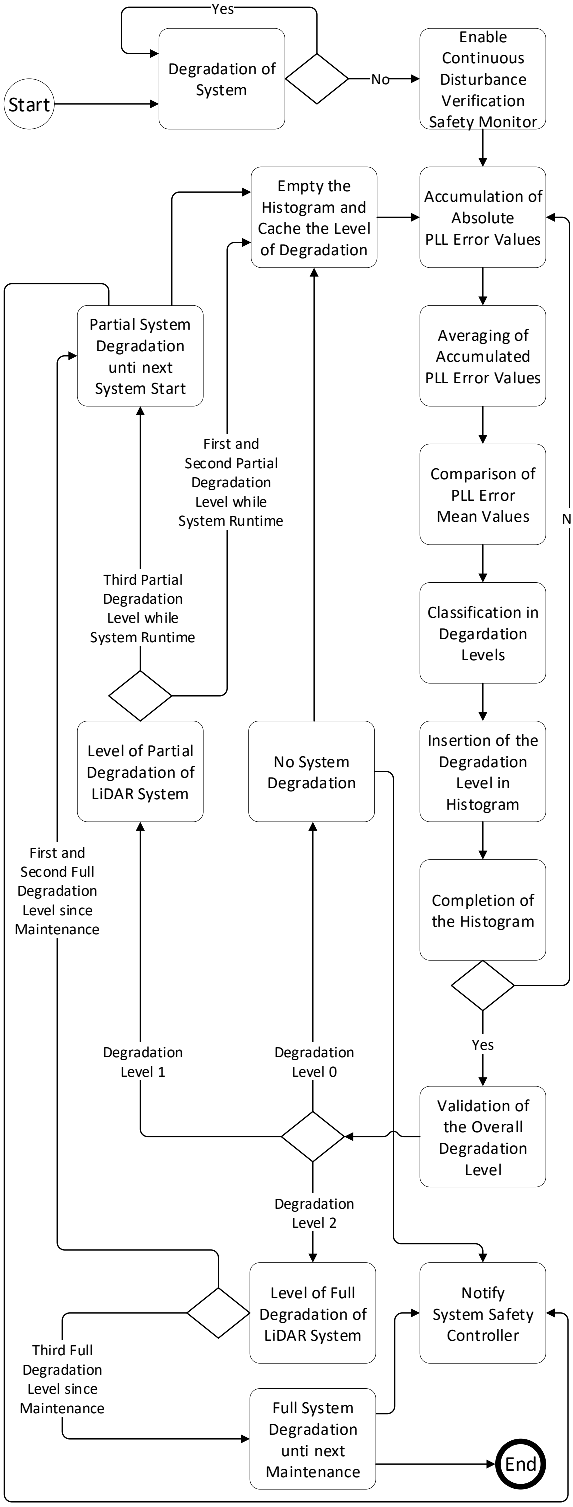

Yes
Degradation of System
No
Enable Continuous Disturbance Verification Safety Monitor
Start
Empty the Histogram and Cache the Level of Degradation
Accumulation of Absolute PLL Error Values
Partial System Degradation unti next System Start
Averaging of Accumulated PLL Error Values
First and Second Partial Degradation Level while System Runtime
Comparison of PLL Error Mean Values
No
Third Partial Degradation Level while System Runtime
Classification in Degardation Levels
Level of Partial Degradation of LiDAR System
No System Degradation
Insertion of the Degradation Level in Histogram
First and Second Full Degradation Level since Maintenance
Completion of the Histogram
Yes
Degradation Level 1
Degradation Level 0
Validation of the Overall Degradation Level
Degradation Level 2
Level of Full Degradation of LiDAR System
Notify System Safety Controller
Third Full Degradation Level since Maintenance
Full System Degradation unti next Maintenance
End

Figure 9. Process flow of the Continuous Disturbance Verification Safety Monitor module.

with the Degradation Levels of the PEMVs until the specified number of histogram entries is reached. Once the histogram is filled, the validation of the Overall Degradation Level is performed.

7) **Validation of Overall Degradation Level**
The validation of the Overall Degradation Level is done by evaluating the individual classes of the Degradation Levels in the histogram. The class that has the largest amount is selected as the Overall Degradation Level. Depending on the resulting Degradation Level, the further steps can be taken. In principle it leads to one of the following actions:

   a) **Degradation Level 0**
   In case the validation results in a Degradation Level 0, then the system is considered reliable and will not be degraded. The histogram is cleared and the Degradation Level is cached and reported to the System Safety Controller via the LSDI. Afterwards the monitoring process restarts with accumulation of absolute PLL error values.

   b) **Degradation Level 1**
   If the monitoring process leads to a Degradation Level 1, then there is not necessarily a system degradation. Until the 3rd time, the system is treated as at Degradation Level 0. Therefore, the histogram is cleared and the level of degradation is cached. But unlike Degradation Level 0, the System Safety Controller is not informed about a new level of degradation. However, if it happens for the 3rd time during system runtime that a Degradation Level 1 results, the system will be partially degraded. The System Safety Controller will be informed via the LSDI and gets a lower priority in the sensor fusion of the environment perception system. Afterwards, the histogram is cleared and the Degradation Level is cached, just like before. The monitoring process starts again.

   c) **Degradation Level 2**
   Should a Degradation Level 2 occur during the monitoring process, there are two possibilities, similar to the Degradation Level 1. For the first two occurrences of Degradation Level 2 after a performed maintenance the system is partially degraded. Here, it is the same as for the 3rd time of Degradation Level 1. The systems priority in sensor fusion is downgraded, until the next system restart and the System Safety Controller is informed via the LSDI. Then the histogram is cleared again and the level of degradation is cached. The monitoring process starts again. However, if there is a 3rd occurrence of Degradation Level 2 since maintenance, the system is completely degraded and the System Safety Controller is informed via the LSDI. The system remains degraded until the next maintenance. The system degradation can then only be removed by an authorized technician or mechanic after the system has been inspected and, if necessary, repaired respectively broken components replaced.

This new monitoring procedure creates another possibility for early detection and reaction to disturbances in MEMS-based LiDAR systems. The system can then be degraded in order to avoid transmitting any erroneous data to the environment perception system. This procedure can help detection of imminent MEMS mirror failures due to aging or MEMS mirror fractures caused by massive shocks and to early initiate required maintenance. Section IV discusses and explains the results of the novel CodeIso approach.

IV. Results

In this section, we provide the measurement results and analysis of our novel monitoring procedures, which have been introduced in Section III.

A. *Novel Safety-Critical Mirror Driver Monitor Evaluation*

Figure 10 shows the start of the novel monitor procedure. After 427 Mirror Half Periods, the frequency of the simulated mirror is changed. The Angle_Ok signal can be used as an indicator for a frequency shift between mirror and driver, because it indicates whether the angle setpoint is reached or not. At the beginning of the frequency mismatch, this indication is also clearly visible in the ZC measurement. The red signal corresponds to the ZC reference signal of the MEMS mirror Driver and the blue one to the ZCsim signal. After the 427th Mirror Half Period it is clearly visible that the reference and the simulated ZC signal are no longer synchronous. The exemplary course of the mirror is recorded at Mirror Angle. The red curve indicates the course of the mirror at the same frequency and the blue curve looks like the course when the new frequency is set for the simulated mirror. Figure 11 shows that the frequency of the mirror has been adjusted again and that the angle setpoint has been reached again from the 1709th Mirror Half Period onwards. Here the Angle_Ok signal is essential for detecting whether the angle setpoint has already been reached again. The frequencies of mirror and Driver are equalized before the 1709th Mirror Half Period. The exemplary courses of the mirror overlap almost completely, reference and simulated ZC signal also occur again almost simultaneously.

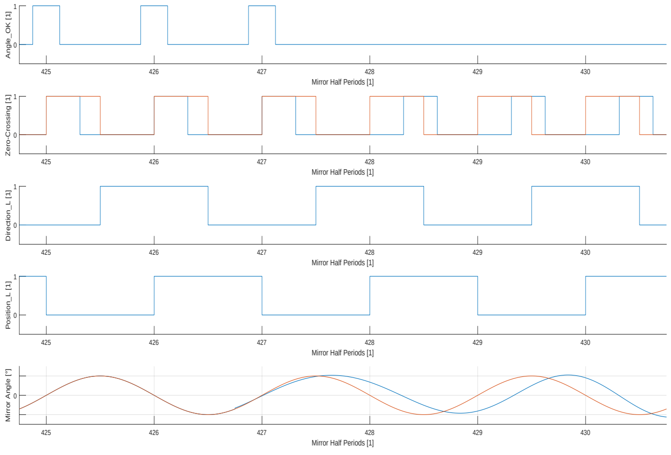


Figure 10. Measurement with the initial frequency adaption of the simulated MEMS mirror.

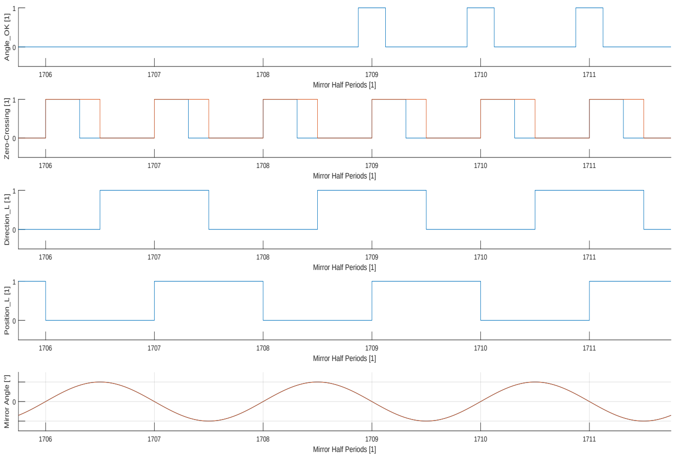


Figure 11. Measurement with the frequency match of the simulated MEMS mirror and the MEMS Driver.

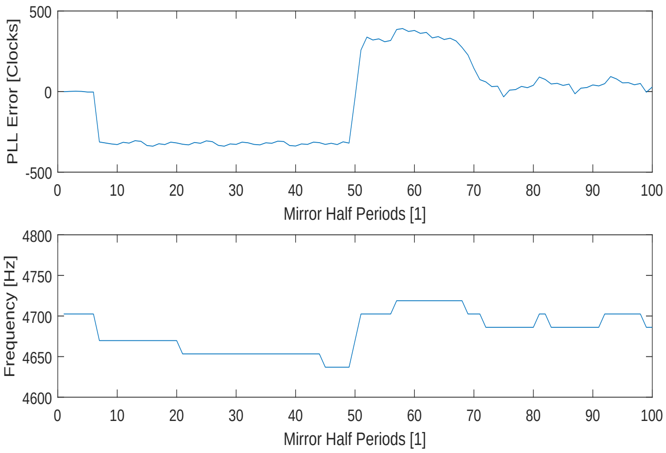


Figure 13. Measurement of the PLL error value accumulation of the CodeIso with an injected massive shock.

For our measurement, the control required 1282 Mirror Half Periods to adjust the frequencies. That was approximately 220 ms for the frequency range from about 2300 Hz to about 2400 Hz. Depending on the frequency difference between mirror and Driver, this control time can be extended or shortened. Finally, the results of the frequency adaption duration are summarized and shown in Table I.

TABLE I. MEASUREMENT RESULTS of SCMDM

	Begin	End	Time in ms
Duration of Frequency Adaption	427	1709	∼ 220

B. Continuous Disturbance Verification Safety Monitor Evaluation

To test the CodeIso, different scenarios were examined and evaluated. The CodeIso accumulates PLL error values of 100 Mirror Half Periods per test run. The first recorded measurement, which is shown in Figure 12, was recorded without any influence. Here one can see that the PLL error value is close to zero and the frequency remains constant. As shown in Table II, the first measurement results in an average of the absolute PLL error values of 3.42. The classification is determined in advance. In our evaluation of the CodeIso, we defined the limits of the different classes exemplarily to show how the division into the specified classes happens. Degradation Level 0 is divided from 0 up to MPEV plus 10, Degradation Level 1 from MPEV plus 10.01 to MPEV plus 50 and Degradation Level 2 from MPEV plus 50.01. Since a reference measurement, which we consider as the maintenance measurement, was calculated to be an average of the absolute PLL error values of 3.38, it is clear that the measurement in Figure 12 belongs to the Degradation Level 0 class. The classification of the different measurements during a CodeIso run is also shown in Table II.

TABLE II. MEASUREMENT RESULTS of CODEISO

Measurement	PLL Error Mean Value	Classified Degradation Level
Maintenance	3.82	
1.	3.42	0
2.	215.93	2
3.	24.23	1
4.	10.98	0
5.	198.58	2
6.	28.04	1
7.	6.09	0
8.	7.53	0
9.	5.32	0
10.	206.45	2

Figure 13 shows the 2nd measurement. Here a massive

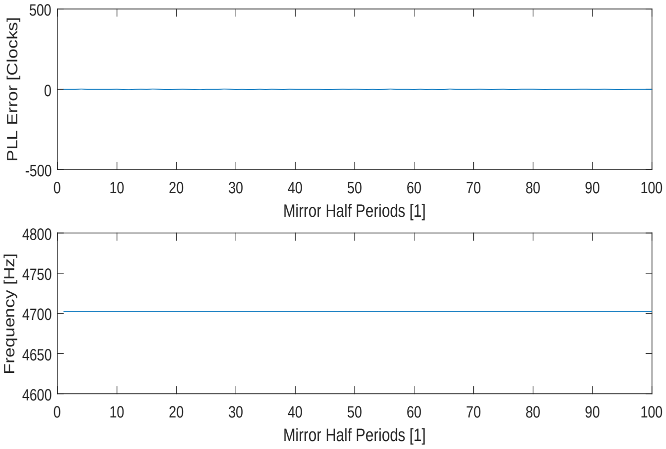


Figure 12. Measurement of the PLL error value accumulation of the CodeIso without any abnormalities.

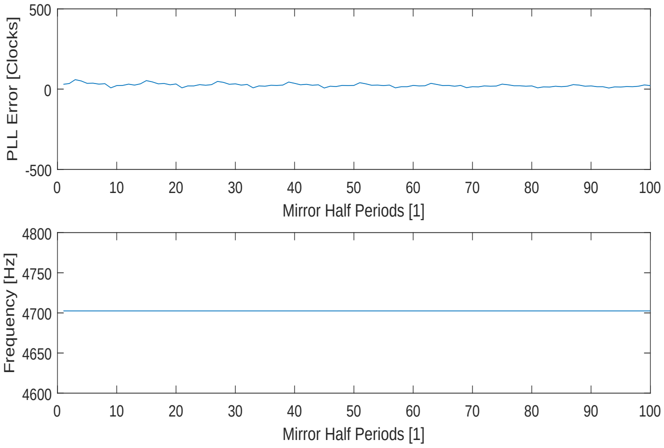


Figure 14. Measurement of the PLL error value accumulation of the CodeIso with effects of the injected massive shock in the measurement before.

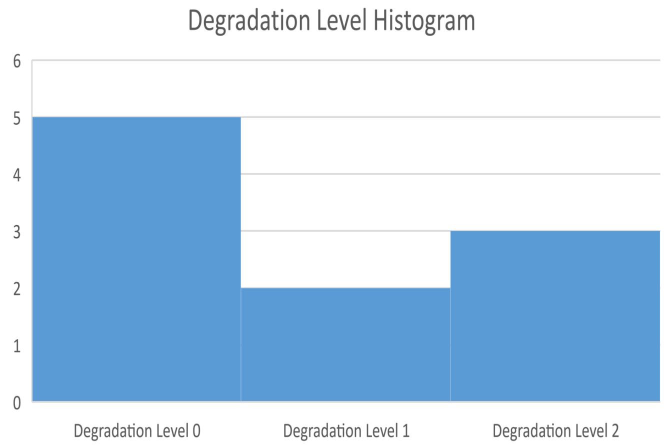


Figure 15. Histogram of a CodeIso iteration with 10 accumulation runs of absolute PLL error values.

shock was injected, simulating a massive shift between MEMS mirror frequency and MEMS Driver frequency. Besides the large PLL error, the unstable frequency clearly shows that the system has been heavily affected. If such a measurement result is obtained, it is clear that it must be noted as Degradation Level 2 in the histogram. As mentioned before, the absolute PLL error value is compared with the MPEV. With 215.93 the PLL error mean value is clearly in the Degradation Level 2 class, because it exceeds MPEV plus 50.01. However, if a majority of such results are obtained, it can be concluded that there are either age-related problems with the MEMS mirror or that the MEMS mirror has been sustainably damaged by a previous massive shock. Furthermore, the 3rd measurement is shown in Figure 14. Here you can see the effects of the previous measurement with the injected massive shock. The frequency is constant again, but the PLL error has not yet settled. Due to the previously defined limits, the 3rd measurement with an absolute PLL error average of 24.23 is slightly in the Degradation Level 1 class. After ten iterations, the histogram shown in Figure 15 is filled with the classified Degradation Levels from Table II. This histogram is now used to validate the Overall Degradation Level. For this purpose, the number of occurrences in the different classes is multiplied by the respective, previously defined factor. A single Degradation Level 2 will not be decisive for the degradation. Depending on the selected factors, more or less of such Degradation Level 2 ratings will be needed to fully degrade the system. The class that contains the highest value will be used as the Degradation Level for the entire LiDAR system. In our case we chose factors 1, 1.5 and 2 for Degradation Level 0, 1 and 2. Figure 16 shows the result for validation. With the highest class value of 6 marked in red, the LiDAR system is set to full degradation. Since we injected a massive shock in three measurements and therefore simulated a heavy damage, respective impairment of the system, it is the result we expected. In case two classes have the same value, the higher Degradation Level is always taken.

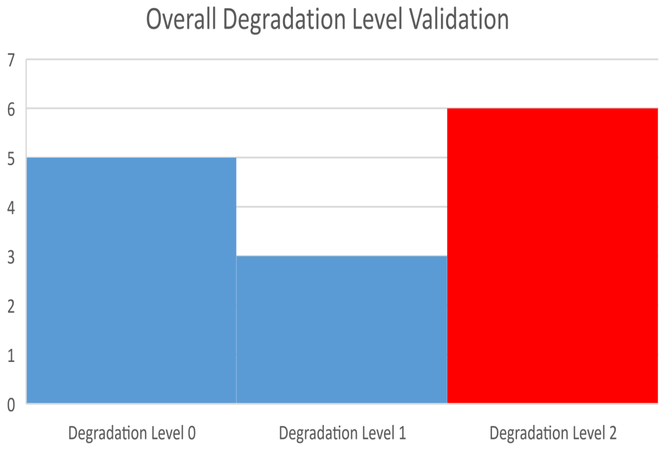


Figure 16. Validation of the Overall Degradation Level by reference to the completed histogram.

V. Conclusion

In our paper, we introduced two novel safety monitor architectures for a Safety-Critical Mirror Driver. With the first monitor we suggest a new possibility to test the control of a MEMS-based LiDAR system and to monitor the functionality of the Driver during the test cycle. The diversity of system monitor options is further increased with this new SCMDM, along with BIST and other diagnostic variants, further reducing the probability of malfunctions remaining undetected. With a duration of around 220ms, this test run is also well under 1s. Therefore, it is unproblematic to perform this procedure during the start of the vehicle or at a vehicle stop in front of a traffic light. Even if the traffic starts to move again, not even 1s passes until the LiDAR system is operational again. Due to the speed at which the vehicle starts to move (usually a slow start), it is only a few centimetres at most that the vehicle does not receive any information from the LiDAR. By further optimizing the parameters, the time required for the test run can probably be shortened considerably. Our intention was to show that in principle it is possible to simulate the mirror and thus create a further possibility for MEMS Driver monitoring by means of the novel monitor. The second monitor we suggest, is a new possibility to continuously check the system for disturbances in the PLL control loop. The CodeIso is used during continuously throughout system operation and is supposed to inform the system of the different Degradation Levels. The absolute PLL error mean values over a given measurement period are used to obtain classified entries in a histogram. After the histogram is filled with the given number of measurements an Overall Degradation Level is determined. In case the MEMS mirror is operated in a frequency range from about 2300 Hz to about 2400 Hz, a statement on the Overall Degradation Level can be made after approximately 10 ms. The weight factors for the Overall Degradation Level were determined exploratory and can also be adapted to get an earlier system degradation or later. Its intention was to design a monitor that detects disturbances in the PLL early and alerts the environment perception system accordingly. With the full degradation of the LiDAR system by this monitor, maintenance of the system becomes necessary. Furthermore, this monitoring procedure extends the diversity of the safety monitors. Monitors as presented here will be even more important in the future for highly automated vehicles than they already are in safety-critical vehicle components. The top priority is to ensure the safety and reliability of the ADAS in the vehicles and also to check whether this is the case.

Acknowledgment

The authors would like to thank all national funding authorities and the ECSEL Joint Undertaking, which funded the PRYSTINE project under the grant agreement number 783190.

PRYSTINE is funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program "ICT of the Future" between May 2018 and April 2021 (grant number 865310). More information: https://iktderzukunft.at/en/.

References

[1] P. Stelzer, A. Strasser, P. Pannagger, C. Steger, and N. Druml, "Monitor for Safety-Critical Mirror Drivers of MEMS Micro-Scanning LiDAR Systems," The Tenth International Conference on Performance, Safety and Robustness in Complex Systems and Applications (PESARO 2020), 2020, pp. 7–12.

[2] N. Druml, G. Macher, M. Stolz, E. Armengaud, D. Watzenig, C. Steger, T. Herndl, A. Eckel, A. Ryabokon, A. Hoess, S. Kumar, G. Dimitrakopoulos, and H. Roedig, "PRYSTINE - PRogrammable sYSTems for INtelligence in AutomobilEs," in 2018 21st Euromicro Conference on Digital System Design (DSD), Aug 2018, pp. 618–626.

[3] European Road Safety Observatory, "Advanced driver assistance systems," https://ec.europa.eu/transport/road_safety/specialist/observatory/analyses/traffic_safety_syntheses/safety_synthesies_en, retrieved: October, 2019. [Online]. Available: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/ersosynthesis2018-adas.pdf

[4] M. Lu, K. Wevers, and R. V. D. Heijden, "Technical Feasibility of Advanced Driver Assistance Systems (ADAS) for Road Traffic Safety," Transportation Planning and Technology, vol. 28, no. 3, 2005, pp. 167–187. [Online]. Available: https://doi.org/10.1080/03081060500120282

[5] SAE, "SAE International Standard J3016 - Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems," SAE International, Standard, January 2014.

[6] C. Brünglinghaus, "Wie das Recht automatisiertes Fahren hemmt," ATZ - Automobiltechnische Zeitschrift, vol. 117, no. 4, Apr 2015, pp. 8–13. [Online]. Available: https://doi.org/10.1007/s35148-015-0039-0

[7] United Nations Conference on Road Traffic, "19 . Convention on Road Traffic," https://treaties.un.org/pages/ViewDetailsIII.aspx?src=TREATY&mtdsg_no=XI-B-19&chapter=11&Temp=mtdsg3&clang=_en, retrieved: October, 2019. [Online]. Available: https://treaties.un.org/pages/ViewDetailsIII.aspx?src=TREATY&mtdsg_no=XI-B-19&chapter=11&Temp=mtdsg3&clang=_en

[8] N. Druml, I. Maksymova, T. Thurner, D. Van Lierop, M. Hennecke, and A. Foroutan, "1D MEMS Micro-Scanning LiDAR," in The Ninth International Conference on Sensor Device Technologies and Applications (SENSORDEVICES 2018), 09 2018.

[9] Velodyne LiDAR, "HDL-64E," 2016.

[10] C. E. Stroud, A designers guide to built-in self-test. Springer Science & Business Media, 2006, vol. 19.

[11] E. J. McCluskey, "Built-In Self-Test Techniques," IEEE Design Test of Computers, vol. 2, no. 2, April 1985, pp. 21–28.

[12] F. Schuldt, F. Saust, B. Lichte, M. Maurer, and S. Scholz, "Effiziente systematische testgenerierung für fahrerassistenzsysteme in virtuellen umgebungen," 2013, retrieved: October, 2019. [Online]. Available: https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00052570

[13] M. Mauritz, F. Howar, and A. Rausch, "Assuring the Safety of Advanced Driver Assistance Systems Through a Combination of Simulation and Runtime Monitoring," in Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications, T. Margaria and B. Steffen, Eds. Cham: Springer International Publishing, 2016, pp. 672–687.

[14] T. Meany, "Functional Safety for Integrated Circuits," July 2018, retrieved: November, 2019. [Online]. Available: https://www.analog.com/en/technical-articles/a54121-functional-safety-for-integrated-circuits.html

[15] K. McCord, Automotive Diagnostic Systems: Understanding OBD I and OBD II. CarTech Inc, 2011.

[16] B. Shinde, D. S. Kore, and D. S. Thipse, "Comparative Study Of On Board Diagnostics Systems-EOBD, OBD-I, OBD-II, IOBD-I and IOBD-II," International Research Journal of Engineering and Technology (IRJET), 2016.

[17] P. Baltusis, "On Board Vehicle Diagnostics," SAE Technical Paper, Tech. Rep., 2004.

[18] T. D. Durbin and J. M. Norbeck, "The effects of repairs on tailpipe emissions for On-Board Diagnostics II-equipped vehicles with the malfunction indicator light illuminated," Journal of the Air & Waste Management Association, vol. 52, no. 9, 2002, pp. 1054–1063.

Statistical Approach to Evaluating Profitability of Stock Markets

Yoshihisa Udagawa
Faculty of Informatics, Tokyo University of Information Sciences
Chiba-city, Chiba, Japan
e-mail: yu207233@rsch.tuis.ac.jp

Abstract - **Candlestick charting is one of the most popular techniques used to predict short-term stock price trends. Despite popularity, there is still no consistent conclusion for the predictability of the technique mainly due to qualitative description of candlestick patterns. This paper proposes a six parameters model that allows us to define both candlestick patterns and price zones where the patterns occur. It is important to grasp buy and sell opportunities for a successful stock trade. Uptrend reversal candlestick patterns are used to find a buy opportunity to enter a trade in a long position. Three exit criteria are proposed to find a sell opportunity to exit a trade for fixing profits or losses. Simulations to estimate profits of markets are performed using historical daily stock data of the US and Asian stock markets with approximately the same parameter values for the six parameters model and the exit criteria in terms of the standard deviation in statistics. Profitability of the proposed stock trade method is statistically examined by linear regression analysis showing that timing to sell stock is significantly related to profits for the three exit criteria. The results of simulations indicate that the US markets are more profitable than Asian markets under the proposed model.**

Keywords - stock price prediction; technical analysis; candlestick patterns; market exit criteria; trailing stop; profit simulation; global market comparison; regression analysis.

I. INTRODUCTION

This paper is an extension of our previous paper on a performance analysis of international stock markets [1]. In the previous paper, we propose a model for finding great opportunities for investors to buy a stock using candlestick chart patterns, and criteria for selling the stock hopefully to keep profits.

In this paper, the following aspects are added to the original work:

(1) Candlestick chart patterns for finding buy opportunities are defined by formulas in terms of successive candlesticks to generalize well-known uptrend reversal candlestick patterns;

(2) A widely-used criterion for finding sell opportunities, named a *trailing stop* [2], is compared with our original criteria in terms of profit;

(3) The period of stock price data used in our experiment is determined based on quarterly and monthly stock price fluctuation analyses.

Forecasting a direction of future stock prices attracts attention of not only financial investors but also researchers on statistics and computer science. Motivation involves to predict the direction of future prices for successful trading and to develop computer system to support it. While many researches on stock price prediction are limited to specific markets, only a few studies are dealing with multiple stock markets.

Dimson, Marsh, and Staunton [3] discuss performances of global markets including emerging markets and developed markets. Though emerging markets have grown to a significant size up to 2007, developed markets, notably the US markets, have outpaced the growth in emerging markets in the 21st century. They conclude that investors should be modest to invest in emerging markets since exchange rate movements are largely affected by inflation that is prevalent in emerging countries.

Ahmad, Ahmed, Vveinhardt and Streimikiene [4] examine Asian stock markets including KSE100 (Pakistan), Nikkei 225 (Japan), KOSPI (South Korea), and BSE (India) in terms of stock return and volatility. The results of statistical analyses lead to a conclusion that volatility is significantly related to return in each market.

To the best of our knowledge, there are few studies that examine profitability of global markets based on simulation using candlestick patterns for estimating buy opportunity and loss stop criteria for finding sell opportunity on daily stock price data.

The contributions of this paper are as follows:

(I) Proposal of a six parameters model to retrieve candlestick patterns that are both similar in price patterns and price zones, i.e., high- or low-price zone in which they occur.

(II) Proposal of three loss stop criteria to exit trade for fixing profits or losses in case of a long position.

(III) Evaluation of profitability of five major markets in the US and Asia using the proposed model for retrieving similar candlestick patterns and the three loss stop criteria through simulations.

The remainder of the paper is organized as follows. Section II reviews related work. Section III gives backgrounds of candlestick patterns. Section IV proposes a model for stock trade using candlestick patterns and exiting criteria from a stock market. Section V presents empirical results on bullish (uptrend) reversal candlestick patterns using five markets' data in the US and Asia. Section VI concludes the paper with our plans for future work.

II. Related Work

There have been a growing number of studies on predicting future price movements of stock markets. In this section, we review previous studies on performances of global markets and predictabilities of candlestick patterns.

A. Studies on Performances of Global Markets

International investing is believed to bring an advantage of better profits from global markets while managing risks better. Dimson, Marsh, and Staunton [3] discuss that emerging markets achieved a higher profit of 11.7% per year than a developed markets' profit of 10.5% from 1950 to 2019. However, because of the global financial crisis in the 21st century, the average profit on US equities has been an annualized 10.6%, while the world average profit excluding the US has been 5.3% in the 21st century. They conclude that investors should be modest to invest in emerging markets because exchange rate movements are largely affected by inflation in emerging countries in addition to questionable capabilities to maintain a fair market.

Ahmad, Ahmed, Vveinhardt, and Streimikiene [4] study Asian stock markets containing KSE100 (Pakistan), Nikkei 225 (Japan), KOSPI (South Korea), Hang Seng (Hong Kong), Shanghai Stock Exchange (China), and BSE (India) in terms of stock returns and volatility. The results show that KOSPI has the highest average annual return of 12.67%, followed by BSE with 11.61%, while KSE 100 has the least return of 9.31%.

B. Studies disapproving of candlestick patterns

As for candlestick patterns in technical analysis [5], several studies [6]-[8] conclude that they are useless based on the experiments using the stock exchange markets' data in the US, Japan, and Thailand.

Horton [6] studies the profitability of 4 pairs of three-day candlestick patterns on 349 stocks that are selected randomly representing all major industry groups. The main conclusion of his study is that these candlestick patterns create no value for trading individual stocks. Marshall, Young, and Cahan [7] find that for a period of 10 days, candlestick charting strategies are not profitable for Dow Jones Industrial's components from 1992 to 2002 and Japanese equity markets from 1975 to 2004. Based on experiments using stock data in the Stock Exchange of Thailand, Tharavanij, Siraprapasiri, and Rajchamaha [8] conclude that any candlestick patterns cannot reliably predict market directions even with filtering by well-known stochastic oscillators [5].

C. Studies approving of candlestick patterns

Other studies conclude that applying a certain candlestick patterns is profitable at least for short-term trade in the US and Asian stock markets [9]-[15].

Caginalp and Laurent [9] study and favorably evaluate the predictive power of eight three-day reversal candlestick patterns on the S&P 500 index during the period of 1992–1996. They propose to define candlestick patterns as a set of inequalities using opening, high, low, and closing prices. These inequalities are taken over by later studies. Goo, Chen, and Chang [10] define 26 candlestick patterns using modified version of inequalities that are proposed by Caginalp and Laurent. They examine these patterns using stock data of Taiwan markets, and conclude that the candlestick trading strategies are valuable for investors.

Chootong and Sornil [11] propose a trading strategy combining price movement patterns, candlestick chart patterns, and trading indicators. A neural network is employed to determine buy and sell signals. Experimental results using stock data of the Stock Exchange of Thailand show that the proposed strategy generally outperforms the traditional trading methods based on technical indicators [5].

One of the obstacles of candlestick charting is the highly subjective nature of candlestick patterns that are defined using words of natural language and illustrations [5]. Tsai and Quan [12] propose an image processing technique to analyze similarities of candlestick charts for stock prediction instead of using numerical inequality formulas. Their experimental results using Dow Jones Industrial Average index show that visual content extraction and similarity matching of candlestick charts are useful for predicting short-term and medium-term stock movements.

Zhu, Atri, and Yegen [13] examine the effectiveness of five different candlestick reversal patterns for predicting short-term stock movements using data of two Chinese stock markets. The results of statistical analysis suggest that the patterns perform well in predicting price trend reversals.

Jamaloodeen, Heinz, and Pollacia [14] statistically examine whether two of the most popular Japanese candlestick patterns, i.e., *Shooting Star* and *Hammer* patterns [5], have predictive significance to forecast a temporary top and bottom using historical data of the S&P 500 index. They define original formula for each pattern using four parameters, i.e., open, high, low, and closing prices. Their findings include the two patterns are highly reliable when using high price for *Shooting Star* and low price for *Hammer* patterns.

Udagawa [15] proposes a dynamic programing method to skip small and noisy candlesticks to improve predictability of candlestick charting. Experimental results show that the proposed method is effective in predicting both uptrend and downtrend.

III. Candlestick Chart Patterns

This section introduces formation of a candlestick chart. Samples of well-known bullish reversal patterns are described. Criticism of candlestick patterns as a method for predicting stock price movements are also mentioned.

A. Formation of Candlestick

A daily candlestick line is formed with the market's opening, high, low, and closing prices of a specific trading day [5]. Figure 1 represents images of typical candlesticks. The candlestick has a wide part, which is called a *body*, showing the range between the open and close prices of that day's trading. If the closing price is above the opening price, then a hollow candlestick is drawn indicating a bullish (rising) candlestick. If the opening price is above the closing price then a filled candlestick is drawn showing a bearish (falling) candlestick.

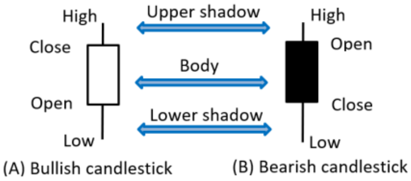


Figure 1. Candlestick formation

The thin lines above and below the body, which are called *shadows*, indicate the high/low ranges. The high price is marked by the top of the upper shadow, and the low price is by the bottom of the lower shadow.

B. Bullish Reversal Candlestick Patterns

Dozens of candlestick patterns are identified and become popular among stock traders [5]. There are three classes of candlestick patterns, i.e., bullish reversal, bearish reversal, and continuation patterns. The reversal patterns are more meaningful because it helps a trader buy at the bottom and sell at the peak of price. This study focuses solely on bullish reversal patterns under the assumption that a trader takes a long position. Triple candlestick patterns are examined because they extend double candlestick patterns with an extra one candlestick for confirmation.

There are four well-known triple candlestick patterns signaling bullish reversal. They are named *morning star*, *three white soldiers*, *three inside up*, and *three outside up*.

Figure 2 shows the *morning star* pattern which is considered as a major reversal signal when it appears in a low-price zone or at a bottom. It consists of three candles, i.e., one short-bodied candle (hollow or filled) between a preceding long filled candlestick and a succeeding long hollow one. The pattern shows that the selling pressure that was there the day before is now subsiding. The third hollow candle overlaps with the body of the first filled candlestick suggests a start of a bullish reversal. The larger the hollow and filled candlesticks are, and the higher the hollow candlestick moves, the stronger the potential reversal.

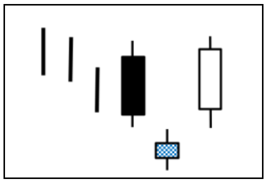


Figure 2. *Morning star* pattern

Figure 3 shows the *three white soldiers* pattern which is interpreted as a strong indication of a bullish market reversal when it appears in a low-rice zone. It consists of three long hollow candlesticks that close progressively higher on each subsequent trading day. Each candlestick opens higher than the previous opening price and closes near the high price of the day, showing a steady advance of buying sentiment.

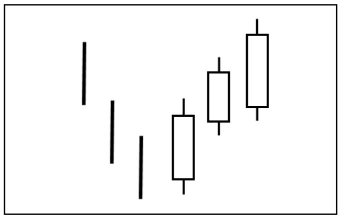


Figure 3. *Three white soldiers* pattern

Figure 4 illustrates the *three inside up* pattern. In this pattern, the first candlestick is a large filled one. The second candlestick is a smaller hollow candlestick contained within the first one. The third candlestick breaks the high price of the second candlestick.

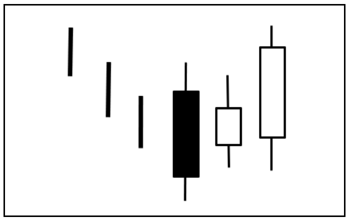


Figure 4. *Three inside up* pattern

Figure 5 illustrates the *three outside up* pattern. It is composed of a small filled candlestick, followed by a longer hollow candlestick that engulfs completely the first one. The third candlestick is a hollow candlestick that closes above the close price of the second one.

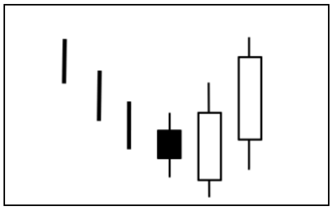


Figure 5. *Three outside up* pattern

In candlestick charting, bullish reversal patterns are deemed to be capable of forecasting price reversal when it appears at bottom after a preceded downtrend. In this study, a price zone where a candlestick occurred is defined by a proposed six parameter model that is described in Section IV.

C. Criticism of Candlestick Patterns

Major criticism of the candlestick chart patterns is that the patterns are qualitatively described with words, such as “long/short candlesticks,” “higher/lower prices,” “strong/weak signal,” supported by some illustrations [5]. Without modeling candlestick patterns in a way that a computer can analyze existence of patterns and perform experiments for measuring a prediction accuracy of future price trends, arguments on the effectiveness of chart patterns would not come to an end.

In addition, some candlestick chart patterns yield different even oppose forecast depending on whether they appear at a high-price or low-price zone. Formulating a suitable definition of price zones is still an open issue.

It deems that because of the lack of the mathematical definition of the candlestick chart patterns, mixed results are obtained in the studies on candlestick charting. Negative conclusions to the predictability of candlesticks are reported [6]-[8], while positive evidences are provided for several candlestick patterns in experiments including the U.S. and the Asian stock markets [9]-[15].

IV. Proposed Model for Stock Trade

This section describes a model to retrieve a candlestick that is similar in both a price change and a price zone where it occurs. Formulas that abstract well-known bullish reversal patterns are defined. Since market exit criteria are vital to keep profits, three criteria including the popular *trailing stop* [2] are proposed.

A. *Six-Parameter Model of Candlestick Retrieval*

After trial and error, we propose a six-parameter model that formalizes a zone where a candlestick occurs in addition to a magnitude of price change and a length of candlestick body. Figure 6 illustrates the proposed model with six parameters defined below:

(1) Change of prices w.r.t previous closing price,
(2) Length of candlestick body,
(3) Difference from 5-day moving average,
(4) Difference from 25-day moving average,
(5) Slope of 5-day moving average,
(6) Slope of 25-day moving average.

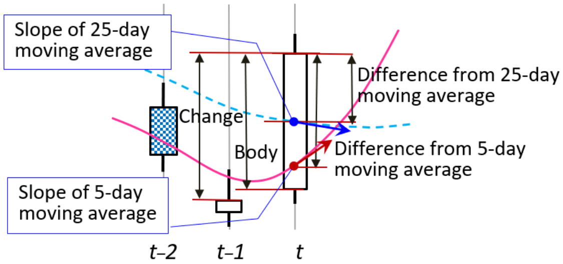

Slope of 25-day moving average
Change
Body
Difference from 25-day moving average
Difference from 5-day moving average
Slope of 5-day moving average
t–2
t–1
t

Figure 6. Six parameters to define candlestick and price zone

While most of the previous studies use a series of inequalities or technical indicators to identify a stock price trend, i.e., an uptrend, downtrend or sideway (a stable range), the proposed model is unique in a sense that it uses two moving averages and their slopes. 5-day and 25-day moving averages are used since they are widely used in Japan. They are significant to identify a zone where a candlestick happens. The slopes of the moving averages are also important to identify the price trend.

Tow candlesticks are defined as similar both in a price and zone if all conditions C_1 to C_6 are satisfied.

C_1: if the difference between a closing price change of a given candlestick and that of a candidate candlestick is within the change tolerance (*change_tol*), then C_1 is true.

C_2: if the difference between a body length of a given candlestick and that of a candidate candlestick is within the body tolerance (*body_tol*), then C_2 is true.

C_3: if the difference between a closing price and a 5-day moving average of a given candlestick and that of a candidate candlestick is within the tolerance (*av5diff_tol*), then C_3 is true.

C_4: if the difference between a closing price and a 25-day moving average of a given candlestick and that of a candidate candlestick is within the tolerance (*av25diff_tol*), then C_4 is true.

C_5: if the slope of a 5-day moving average of a given candlestick and that of a candidate candlestick is within the given tolerance (*slope5_tol*), then C_5 is true.

C_6: if the slope of a 25-day moving average of a given candlestick and that of a candidate candlestick is within the given tolerance (*slope25_tol*), then C_6 is true.

B. *Finding Buy Oppotunities of Stock Trade*

Profit in stock trade in a long position comes from the difference between a buy price and a sell price of a stock. So, buying a stock at a low price and selling it at a higher price is essential for a successful stock trade.

We define formulas that intend to be a generalization of three-day bullish reversal patterns including the *morning star* pattern [5], etc. The formulas in combination with the six parameters model of similar candlesticks are used to find buy opportunity in a low-price zone.

Let CP(t) and OP(t) denote close and open prices of a given market day t. A bullish reversal candlestick pattern is defined as follows:

$$\mathrm{CP}(t) > \mathrm{OP}(t) \qquad (1)$$

$$(\mathrm{CP}(t) + \mathrm{OP}(t)) / 2 > \mathrm{CP}(t\text{–}1) \qquad (2)$$

$$\mathrm{CP}(t) > \mathrm{CP}(t\text{–}2) \qquad (3)$$

Figure 7 depicts a pattern defined by (1)-(3). Inequality (1) means the body of the candlestick is hollow with signaling a rise in stock prices. Inequality (2) specifies that the close price of day *t–1* is below the average of the open and close prices of day *t*. Inequality (3) describes that the close price of day *t–2* is below the close price of day *t*. Inequalities (2) and (3) are satisfied even when the close price of day *t–2* is far below the close price of day *t–1*.

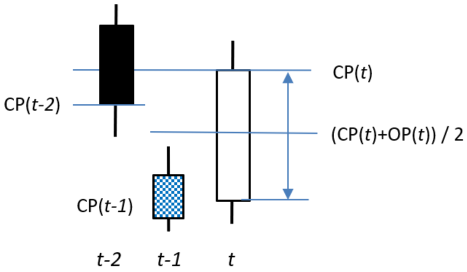


Figure 7. Pattern defined by (1)-(3)

Inequalities (1)-(3) exclude a condition on the length of a candlestick body. In effect, the length of the candlestick body of day *t* is specified by the condition C_2 because (1)-(3) are used with the proposed six parameters model. Since there is no specification on other candlesticks bodies of day *t–1* and *t–2*, (1)-(3) generalize four bullish reversal patterns.

C. Finding Selling Oppotunities of Stock Trade

To successfully complete stock trade, we need to find preferable opportunities to sell a stock in a long position. Candlestick patterns claim that they can be applied to find sell opportunities. However, because there are tens of candlestick downtrend patterns known so far, it is difficult to implement all the patterns.

A capable method named a *trailing stop* [2] is proposed to decide when to sell a stock. The *trailing stop* criteria is designed to lock in profits and suppress losses. Figure 8 illustrates a concept of the criteria. A trader typically specifies a stop price by means of setting a percentage of a loss that can be tolerable on a trade. If a stock price rises in trader's favor, the stop price is continuously reset to a higher value. In case a stock price falls against trader's expectation and exceeds the tolerable percentage of a loss, then the *trailing stop* criterion signals selling a stock.

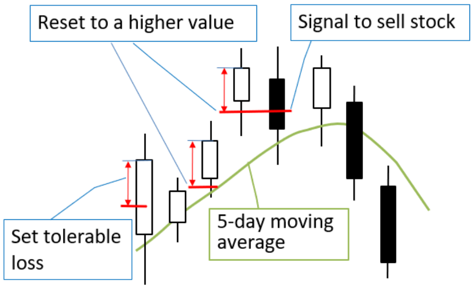


Figure 8. Concept of *trailing stop* criterion

To compare performance of the *trailing stop* criterion with others, we implement two original criteria named a sum of negative price change criterion *(SumNC)*, and a sum of negative price changes below a 5-day moving average (*SumNC5av)*. Their concepts are depicted in Figure 9.

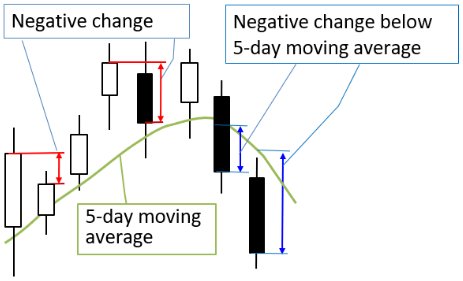


Figure 9. Concept of *SumNC* and *SumNC5av* criteria

The *SumNC* criterion signals selling a stock when the sum of negative price changes exceeds a pre-defined tolerable value. This criterion works the same way as a stop-loss when a price of a stock continues to decline contrary to trader's expectation.

Moving averages are often used in a trading strategy, especially over 5, 25, and 75-day periods in Japan. The *SumNC5av* criterion is devised as a criterion of stock trading with respect to a moving average. The criterion keeps holding a stock until the sum of the negative differences between a stock price and a 5-day average reaches below a pre-defined value. Because falls of a stock price often keep above a 5-day price average in an uptrend, e.g., the fourth candlestick from the right in Figure 9, the *SumNC5av* criterion tends to hold a stock longer than the *SumNC* criterion.

V. Empirical Results

After outlining processes of the performed experiments, statistical analyses of price fluctuations on Dow Jones Industrial Average, NASDAQ Composite index, Nikkei 225 Stock Average, Hang Seng index, and Shanghai Composite index are presented. Results of profit simulations using historical daily stock data of five stock markets are discussed.

A. Data Conversion

Stock prices are converted to the ratio of closing prices to reduce the effects of highness or lowness of the stock prices. The formula below is used for calculating the ratio of prices as a percentage.

$$R_i = (CP_i - CP_{i-1})*100 / CP_i \quad (1 \leqq i \leqq n) \qquad (4)$$

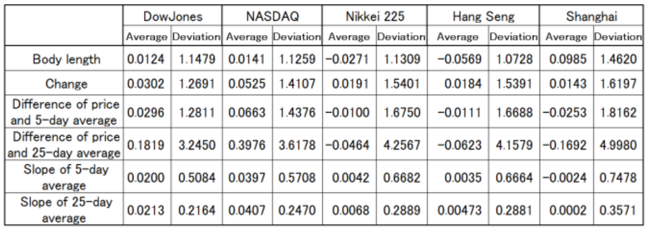
CP_i indicates the closing price of the i-th market date. CP_n means the closing price of the current date. R_n is the ratio of the difference between the closing price CP_n of the current date and the closing price CP_{n-1} of one day before to the CP_n.

The daily stock data from Mar. 1, 2007 to June 30, 2020 are used in the experiment. The number of data is approximately 3,358 for each market. Daily stock data are downloaded from a website that provides historical data of major world markets [16].

B. Statistics of Candlestick Parameters

As the first step of experiments, quarterly statistics about six parameters concerning the proposed six-parameter model of a candlestick pattern are calculated for the period between Apr. 1, 2007 and Jun. 30, 2020. Table I summarizes statistics of the six parameters for the five markets, i.e., Dow Jones, NASDAQ, Nikkei 225, Hang Seng, and Shanghai.

TABLE I. SUMMARY OF STATISTICS OF SIX PARAMETERS DURING PERIOD BETWEEN APR. 1, 2007 AND JUN. 30, 2020



	DowJones		NASDAQ		Nikkei 225		Hang Seng		Shanghai	
	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation
Body length	0.0124	1.1479	0.0141	1.1259	−0.0271	1.1309	−0.0569	1.0728	0.0985	1.4620
Change	0.0302	1.2691	0.0525	1.4107	0.0191	1.5401	0.0184	1.5391	0.0143	1.6197
Difference of price and 5-day average	0.0296	1.2811	0.0663	1.4376	−0.0100	1.6750	−0.0111	1.6688	−0.0253	1.8162
Difference of price and 25-day average	0.1819	3.2450	0.3976	3.6178	−0.0464	4.2567	−0.0623	4.1579	−0.1692	4.9980
Slope of 5-day average	0.0200	0.5084	0.0397	0.5708	0.0042	0.6682	0.0035	0.6664	−0.0024	0.7478
Slope of 25-day average	0.0213	0.2164	0.0407	0.2470	0.0068	0.2889	0.00473	0.2881	0.0002	0.3571

Averages of all six parameters are positive for Dow Jones and NASDAQ indicating that the two markets are generally on an uptrend. Nikkei 225 and Han Seng markets mark negative values for three parameters, i.e., the candlestick body length, the difference between price and 5-day average, and the difference between price and 25-day average. Shanghai market has negative values of three parameters, i.e., the difference between price and 5-day average, the difference between price and 25-day average, and the slope of 5-day average. These negative values suggest that the Asian markets are less profitable than the US ones.

Figure 10 shows price fluctuations as a percentage of the five markets for each quarter. “200706” in the x-axis of Figure 10 indicates the second or April-June quarter of calendar year 2007, for example. We see that the prices of Shanghai and Hang Seng markets fluctuate larger than those of the other markets, notably during the period from the second quarter of 2007 to the fourth quarter of 2015.

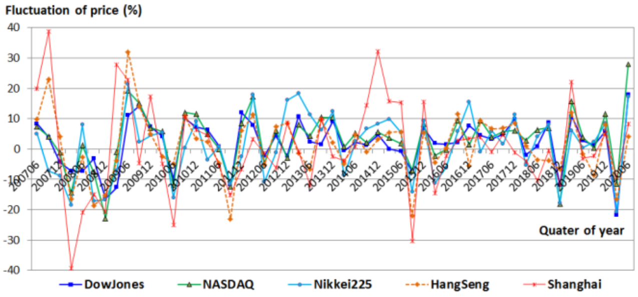


Figure 10. Price fluctuations of five markets for each quarter

We take the period between Apr. 1, 2015 and Jun. 30, 2020 for further examination, because the price fluctuations of the five markets are somehow linked during this period as observed in Figure 10. Table II summarizes monthly statistical results of the six parameters. The average values of all six parameters are positive for Dow Jones and NASDAQ markets. All average values are barely positive for Nikkei 225 market. The five average values out of six parameters are negative for Han Seng and Shanghai markets.

TABLE II. SUMMARY OF MONTHLY STATISTICS OF SIX PARAMETERS DURING PERIOD BETWEEN APR. 1, 2015 AND JUN. 30, 2020

	DowJones		NASDAQ		Nikkei 225		Hang Seng		Shanghai	
	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation
Body length	0.0046	0.9296	0.0236	0.9684	−0.0195	0.9551	−0.0499	0.8753	0.1086	1.2899
Change	0.0361	1.2520	0.0630	1.3101	0.0201	1.3242	0.0057	1.2012	−0.0069	1.4673
Difference of price and 5−day average	0.0412	1.2807	0.0922	1.3172	0.0039	1.4657	−0.0166	1.3327	−0.0588	1.6885
Difference of price and 25−day average	0.2537	3.4524	0.5442	3.4125	0.0370	3.7052	−0.0839	3.4422	−0.3279	4.5487
Slope of 5−day average	0.0261	0.5084	0.0521	0.5252	0.0088	0.5962	−0.0017	0.5474	−0.0198	0.6864
Slope of 25−day average	0.0277	0.2261	0.0514	0.2321	0.0117	0.2505	−0.00060	0.2394	−0.0137	0.3159

Figure 11 shows price fluctuations of the five markets for each month. For example, “202006” in the x-axis of Figure 11 indicates Jun. 2020.

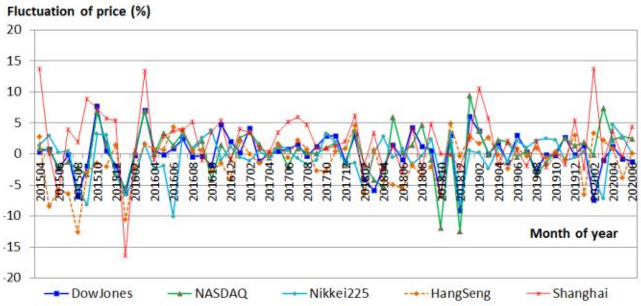
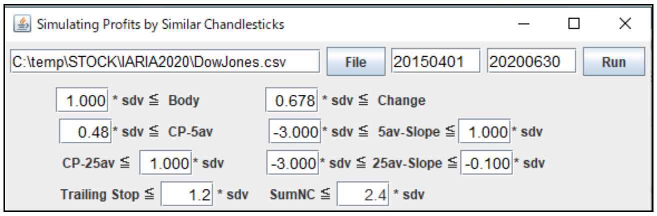


Figure 11. Price fluctuations of five markets for each month

In Figure 11, it is observed that stock price fluctuations of Dow Jones and NASDAQ overlap in many months. Stock prices of Asian markets roughly move in the same direction with some degrees of time delays.

C. GUI for Experiments

Figure 12 shows a GUI that is used in the experiments. It provides parameter values for the six-parameter model in Figure 6 and the three market-exit criteria. The *File* button allows users to choose a CSV file containing a set of stock price data. The full path of the CSV file is displayed. In Figure 12, a file named *DowJones.csv* is chosen. The right two text boxes in the first-row are used to specify periods of market days, i.e., 20150401 to 20200630, used in the experiments.



Simulating Profits by Similar Chandlesticks

C:\temp\STOCK\IARIA2020\DowJones.csv | File | 20150401 | 20200630 | Run

1.000 * sdv ≦ Body   0.678 * sdv ≦ Change

0.48 * sdv ≦ CP-5av   -3.000 * sdv ≦ 5av-Slope ≦ 1.000 * sdv

CP-25av ≦ 1.000 * sdv   -3.000 * sdv ≦ 25av-Slope ≦ -0.100 * sdv

Trailing Stop ≦ 1.2 * sdv   SumNC ≦ 2.4 * sdv

Figure 12. GUI used in experiments

A developed simulator calculates the averages and the standard deviations of the six parameters for the specified period. The text box in the second-row labeled by *“* std ≦ Body”* specifies the magnitude of standard deviation that the length of a candlestick body need to satisfy. Because the experiments are performed for a long position, the length of a candlestick body is required to be longer than the specified magnitude of the standard deviation. The text box labeled by *“* std ≦ Change”* specifies the magnitude of standard deviation that price changes need to satisfy.

Three text boxes in the third-row are used to define ranges of the difference between a close price and a 5-day moving average (labeled by *“CP-5av”*), and a slope of a 5-day moving average (labeled by *“5av-Slope”*). In order to spot candlesticks in uptrend reversal, the developed simulator is designed to retrieve candlesticks whose *CP-5av* are greater than the specified magnitude. As for *5av-Slope*, we need to specify both lower and upper limits.

Three text boxes in the fourth-row are used to define ranges of the parameters concerning 25-day average. The three text boxes in the fourth-row play the similar role as those in the third-row.

Tow text boxes labeled by "*Trailing Stop"* and "*SumNC"* at the bottom of the GUI specify parameter values for the *Trailing Stop* and *SumNC* criteria. The value labeled by "*SumNC"* is also applied to the *SumNC5av* criterion.

Trade price is calculated by the average of the high and low prices on a trading day. This calculation is feasible because the high and low prices can be observed during stock trading time. Traders can decide whether to keep or sell a stock based on the prices. Simulated profits are calculated using the trade price, i.e., the average of the high and low prices.

A typical commission fee of online brokers is between 0.05% and 0.15% depending on the order size of trade. Because we assume swing trading [17] that attempts to capture profit over a period of a few days to several weeks, a commission fee is treated as negligible costs in this study.

D. Experiments on Profit Estimation by Simulation

Table III shows an experimental result performed on Dow Jones daily data using parameters shown in Figure 12. The column named "*Trade day"* in Table III lists market days that satisfy all conditions defined by (1) to (3), and C_1 to C_6. Strictly, conditions defined by (2) and (3) are embedded in source code and cannot be seen on the GUI.

The *trailing stop*, *SumNC*, and *SumNC5av* criteria are used to make a decision to sell a stock. The parameter value for the *trailing stop* is set to 1.2 times the standard deviation of price changes. Because the standard deviation of price changes is 1.2520% as shown in Table II, the tolerable loss for the *trailing stop* is 1.5024% (=1.2*1.2520%).

The parameter values for the *SumNC* and *SumNC5av* are set to 2.4 times the standard deviation of price changes. The tolerable loss is analogously calculated to be 3.0048% (=2.4*1.2520%). The other parameters are carefully adjusted for each market to retrieve 26 days for buy opportunities so that the number of the days is suitable for being analyzed based on the theory of the normal distribution [18].

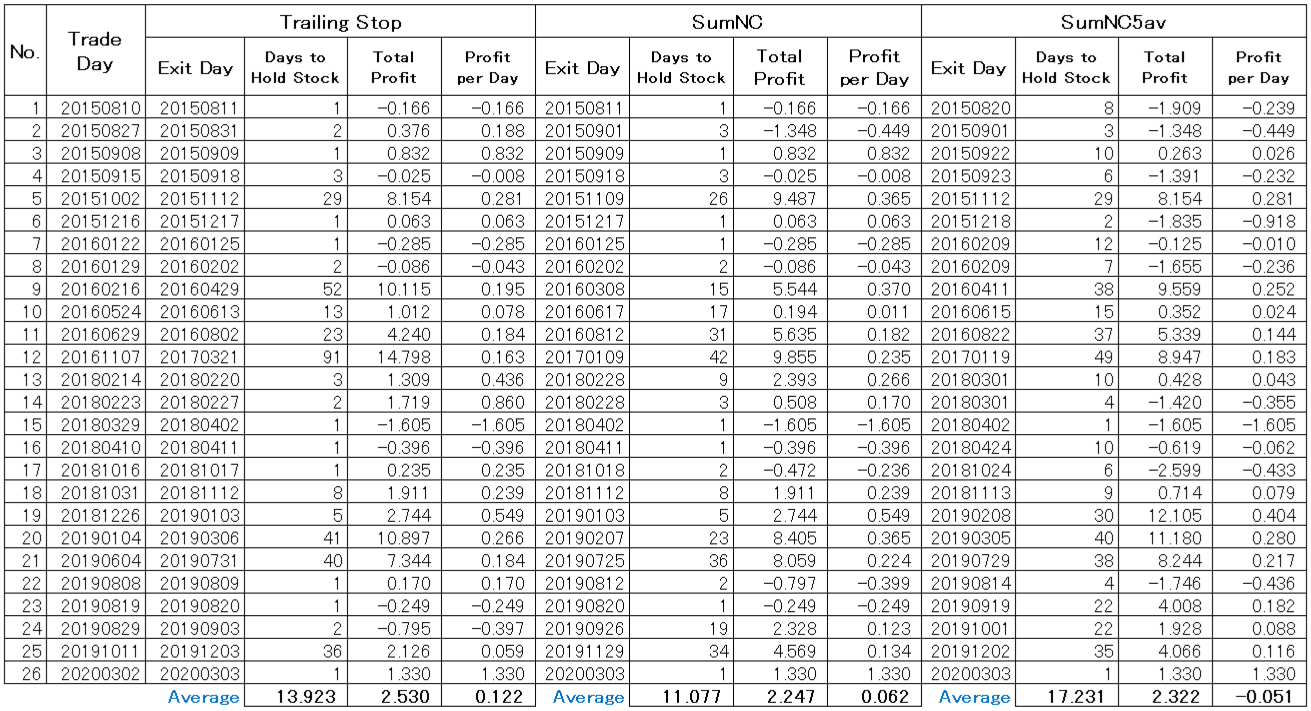
The columns named "*Exit day"*, *"Days to hold stock"*, "*Total profit"*, and "*Profit per day"* indicate the day to sell stocks, the number of holding days of a stock, a simulated total profit, and a rate of a profit to the number of holding days of a stock, respectively. Averages of profits are 2.53%, 2.247%, and 2.322% for the *trailing stop*, *SumNC*, and *SumNC5av* criteria, respectively.

Figure 13 shows a candlestick chart of the trade that buy a stock on Feb. 16, 2016 as listed in the 9th line of Table III. The *trailing stop*, *SumNC*, and *SumNC5av* criteria generate a signal to close the trade after 52, 15, and 38 days, respectively.

The *trailing stop* criterion signals selling a stock on Apr. 29, 2016, i.e., 52 days after the stock trade. Because the nature of the *trailing stop* criterion, there is always a predefined amount of loss from the maximum profit before closing a trade, i.e., 1.5024% in this experiment.

As for the *SumNC* criterion, the sum of negative price changes exceeds the predefined limit value of 3.0048% on Mar. 8, 2016, i.e., 15 days after the stock trade.

TABLE III. ESTIMATED PROFITS ON DOW JONES DAILY DATA



No.	Trade Day	Trailing Stop				SumNC				SumNC5av			
		Exit Day	Days to Hold Stock	Total Profit	Profit per Day	Exit Day	Days to Hold Stock	Total Profit	Profit per Day	Exit Day	Days to Hold Stock	Total Profit	Profit per Day
1	20150810	20150811	1	−0.166	−0.166	20150811	1	−0.166	−0.166	20150820	8	−1.909	−0.239
2	20150827	20150831	2	0.376	0.188	20150901	3	−1.348	−0.449	20150901	3	−1.348	−0.449
3	20150908	20150909	1	0.832	0.832	20150909	1	0.832	0.832	20150922	10	0.263	0.026
4	20150915	20150918	3	−0.025	−0.008	20150918	3	−0.025	−0.008	20150923	6	−1.391	−0.232
5	20151002	20151112	29	8.154	0.281	20151109	26	9.487	0.365	20151112	29	8.154	0.281
6	20151216	20151217	1	0.063	0.063	20151217	1	0.063	0.063	20151218	2	−1.835	−0.918
7	20160122	20160125	1	−0.285	−0.285	20160125	1	−0.285	−0.285	20160209	12	−0.125	−0.010
8	20160129	20160202	2	−0.086	−0.043	20160202	2	−0.086	−0.043	20160209	7	−1.655	−0.236
9	20160216	20160429	52	10.115	0.195	20160308	15	5.544	0.370	20160411	38	9.559	0.252
10	20160524	20160613	13	1.012	0.078	20160617	17	0.194	0.011	20160615	15	0.352	0.024
11	20160629	20160802	23	4.240	0.184	20160812	31	5.635	0.182	20160822	37	5.339	0.144
12	20161107	20170321	91	14.798	0.163	20170109	42	9.855	0.235	20170119	49	8.947	0.183
13	20180214	20180220	3	1.309	0.436	20180228	9	2.393	0.266	20180301	10	0.428	0.043
14	20180223	20180227	2	1.719	0.860	20180228	3	0.508	0.170	20180301	4	−1.420	−0.355
15	20180329	20180402	1	−1.605	−1.605	20180402	1	−1.605	−1.605	20180402	1	−1.605	−1.605
16	20180410	20180411	1	−0.396	−0.396	20180411	1	−0.396	−0.396	20180424	10	−0.619	−0.062
17	20181016	20181017	1	0.235	0.235	20181018	2	−0.472	−0.236	20181024	6	−2.599	−0.433
18	20181031	20181112	8	1.911	0.239	20181112	8	1.911	0.239	20181113	9	0.714	0.079
19	20181226	20190103	5	2.744	0.549	20190103	5	2.744	0.549	20190208	30	12.105	0.404
20	20190104	20190306	41	10.897	0.266	20190207	23	8.405	0.365	20190305	40	11.180	0.280
21	20190604	20190731	40	7.344	0.184	20190725	36	8.059	0.224	20190729	38	8.244	0.217
22	20190808	20190809	1	0.170	0.170	20190812	2	−0.797	−0.399	20190814	4	−1.746	−0.436
23	20190819	20190820	1	−0.249	−0.249	20190820	1	−0.249	−0.249	20190919	22	4.008	0.182
24	20190829	20190903	2	−0.795	−0.397	20190926	19	2.328	0.123	20191001	22	1.928	0.088
25	20191011	20191203	36	2.126	0.059	20191129	34	4.569	0.134	20191202	35	4.066	0.116
26	20200302	20200303	1	1.330	1.330	20200303	1	1.330	1.330	20200303	1	1.330	1.330
		Average	13.923	2.530	0.122	Average	11.077	2.247	0.062	Average	17.231	2.322	−0.051

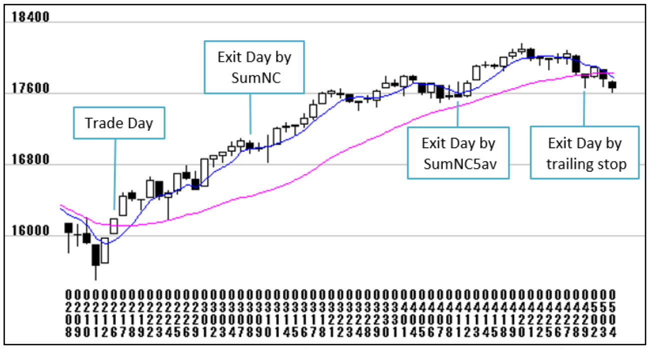


Figure 13. Candlestick chart of trade on Feb. 16, 2016 and after trade

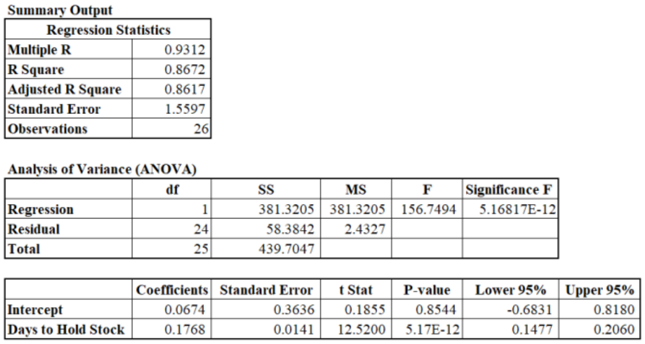
The *SumNC5av* criterion, accumulating negative changes below 5-day moving average, reaches the limit value of 3.0048% on Apr. 11, 2016, i.e., 38 days after the stock trade. Because the stock prices generally keep a steady uptrend after the trade day, the period to hold a stock based on the *SumNC5av* criterion is long more than twice of that of the *SumNC* criterion.

E. Regression Analysis

Regression analysis is a reliable mathematical method to estimate the relationship between two or more variables of interest. It is widely used to examine the influence of one or more independent variables on a dependent variable. In this section, we evaluate profitability and diversity of the proposed stock trade method using regression analysis.

Table IV summarizes the result of regression analysis that is applied to the results for the *trailing stop* criterion by specifying *Profit* as an independent variable and *Days to hold stock* as a dependent variable. *R Square* is 0.8672 suggests that 86.72% of *Profit* can be explained by *Days to hold stock*. The table *ANOVA* (analysis of variance) shows the results of the F-test for measuring the probability that *Profit* is related to *Days to hold stock* by chance. As the value of *Significance F* and *P-value* are 5.16817E-12 (<0.05), which means that *Profit* is significantly related to *Days to hold stock.*

TABLE IV. SUMMARY OF REGRESSION ANALYSIS



Summary Output

Regression Statistics	
Multiple R	0.9312
R Square	0.8672
Adjusted R Square	0.8617
Standard Error	1.5597
Observations	26

Analysis of Variance (ANOVA)

	df	SS	MS	F	Significance F
Regression	1	381.3205	381.3205	156.7494	5.16817E-12
Residual	24	58.3842	2.4327		
Total	25	439.7047			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.0674	0.3636	0.1855	0.8544	-0.6831	0.8180
Days to Hold Stock	0.1768	0.0141	12.5200	5.17E-12	0.1477	0.2060

Figure 14 presents a scatter plot for the *trailing stop* criterion with *Profit* as the x-axis and the *number of days to hold stock* as the y-axis. Figure 14 depicts a significant correlation between the two variables. 16 out of 26 trades are terminated within less than four days. Because of early decision, losses are limited within 1.605% that is shown in the 15th line of Table III. On the other hand, when price moves in a favorable direction, the *trailing stop* criterion leads to hold stock for a rather long period resulting in high profits up to 14.798% that is shown in the 12th line of Table III.

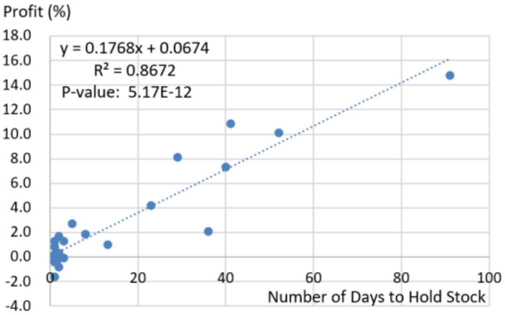


Figure 14. Scatter chart for the *trailing stop* criterion

Figure 15 shows a scatter plot for the *SumNC* criterion. *R Square* is 0.7677. *P-value* is 4.4816E-09 (<0.05). 14 out of 26 trades are stopped within less than four days. Because selling stock is performed based on the sum of the negative prices, the *SumNC* criterion is apt to stop trade earlier than the *trailing stop* criterion.

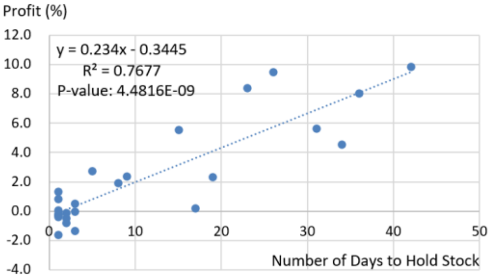


Figure 15. Scatter chart for the *SumNC* criterion

Figure 16 is a scatter plot for the *SumNC5av* criterion. *R Square* is 0.8068. *P-value* is 4.82E-10 (<0.05). In the *SumNC5av* criterion, a sell decision is made using the sum of negative differences between a stock price and a 5-day average. Since a negative price change is ignored while the price keeps above the 5-day average, the *SumNC5av* criterion is insensitive to price fluctuations. Accordingly, 4 out of 26 trades are stopped within less than four days.

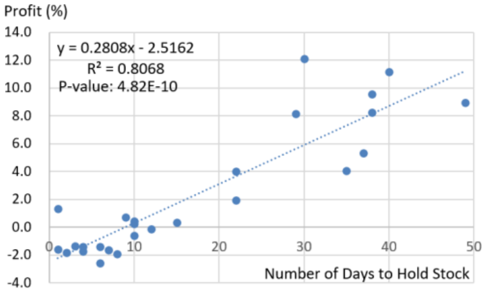


Figure 16. Scatter chart for the *SumNC5av* criterion

All three criteria show significant correlations between the two variables, i.e., *Profit* and *Days to hold stock*, with *R Squares* are greater than 0.7677.

F. Profits for Each Market

Figure 17 shows a graph on the *trailing Stop* exit criterion with profits of each market in the y-axis and multiples of the standard deviation of price changes in the x-axis. In NASDAQ market, the profits increase as the multiples of the standard deviation increase. In Dow Jones market, by contrast, profit reaches a peak of profits at 1.2 multiples of the standard deviation of price changes. Hang Seng and Shanghai markets reach peaks at 1.0 multiples of the standard deviation. Hang Seng and Nikkei 225 markets are notably less profitable than the others.

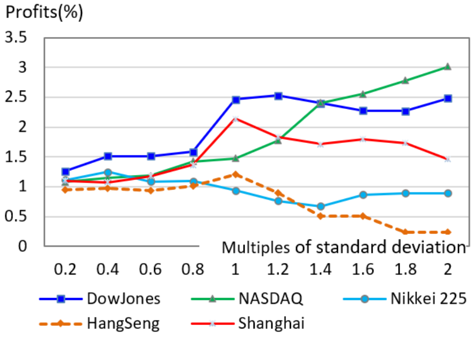


Figure 17. Profits and times of standard deviation using *Trailing Stop*

Figure 18 shows a graph concerning the *SumNC* exit criterion with profits in the y-axis and multiples of standard deviation of price changes in the x-axis. In NASDAQ market, the profits increase as the multiples of the standard deviation increase. In the other four markets, profits reach the highest points at a certain multiple of the standard deviation of price changes. In Dow Jones market, for example, profit reaches a peak at 2.8 multiples of the standard deviation of price changes.

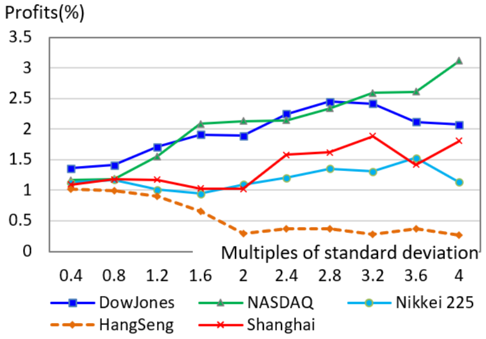


Figure 18. Profits and times of standard deviation using *SumNC*

Figure 19 shows a graph on the *SumNC5av* exit criterion. Profits of Hang Seng market show less than those of the other markets with losses at three multiples, i.e., 1.6, 2.0, and 4.0.

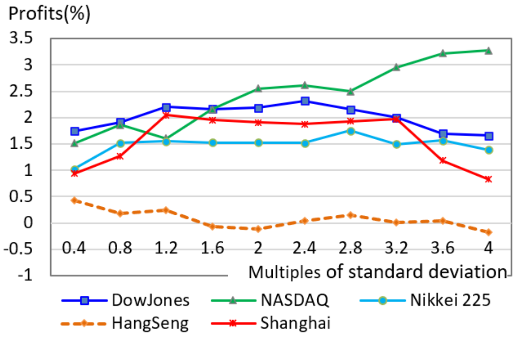


Figure 19. Profits and times of standard deviation using *SumNC5av*

For all three exit criteria, profits of the NASDAC market continue increasing in the range of multiples examined in the experiments. The fact suggests that it is a better decision to hold a stock even if prices fall. Because NASDAQ's stock prices generally keep rising, they tend to recover in a short period. Meanwhile, profits of Dow Jones market always show higher than those of Shanghai, Nikkei 225, and Hang Seng markets, implying that Dow is a leading index of Asian markets.

G. Holding Days for Each Market

Figure 20 shows a graph concerning the *Trailing stop* exit criterion with the number of days to hold a stock in the y-axis and multiples of standard deviation of price changes in the x-axis. The criterion generates a signal to sell a stock when a stock price falls more than a predefined percentage from the highest price. So, the longer the days to hold a stock become, the fewer chances of the stock price plummets are expected. Figure 20 suggests that Dow Jones market has fewer plunges than the other markets.

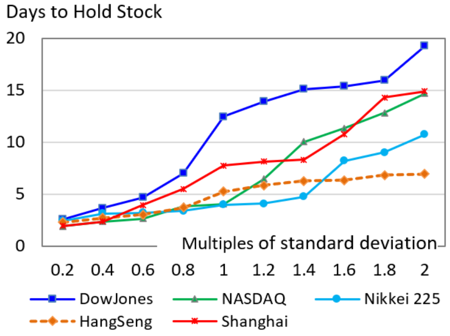


Figure 20. Days to hold stock and times of standard deviation using *Trailing Stop*

Figure 21 is a graph on the *SumNC* exit criterion showing a relationship between the number of days to hold a stock and multiples of the standard deviation of price changes. The number of days to hold a stock apparently linearly depends on multiples of the standard deviation. Since the *SumNC* criterion is based on the sum of the negative price changes, the longer days to hold stock mean the smaller chances of negative price changes during the periods of trade. Dow Jones and Shanghai markets are more likely to continue price uptrends with less falls in stock price than Hang Seng and Nikkei 225 markets.

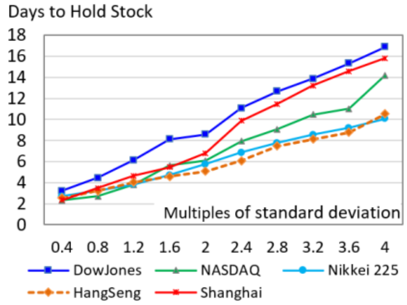


Figure 21. Days to hold stock and times of standard deviation using *SumNC*

Figure 22 shows a graph on the *SumNC5av* exit criterion. Due to the definition of the *SumNC5av* criterion, days to hold stock are tend to be longer than those of the *SumNC* criterion. For example, the maximum number of days to hold in Dow Jones market is 24 for the *SumNC5av* while it is 17 for the *SumNC5* as shown in Figure 21. The graph apparently shows linear dependency between the two variables.

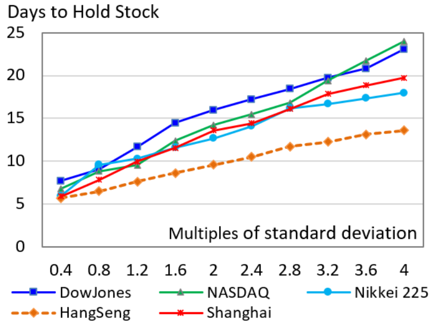


Figure 22. Days to hold stock and times of standard deviation using *SumNC5av*

Days to hold a stock in Dow Jones market tend to be longer than those in the other markets on the three exit criteria, likely leading to high profitability compared with the other markets.

H. Visualizing Profit and Loss Pattern

Figure 23 shows simulated profit-and-loss patterns in a bar graph using the *trailing stop*, *SumNC*, and *SumNC5av* criteria for Dow Jones market. Parameters for the *trailing stop*, *SumNC*, and *SumNC5av* are set to 1.2, 2.4, and 2.4 multiples of the standard deviation, respectively. The profits are sorted in ascending order. Roughly, the three exit criteria result in comparable profits and/or losses. Approximately ten out of 26 trade days result in small amounts of losses with large amounts of profits for the rest of days.

Since the *SumNC5av* criterion tends to hold a stock longer than the *SumNC* criterion, profits and losses obtained by the *SumNC5av* criterion are greater than those by the *SumNC*, suggesting that return and risk are always correlated. Profits gaind by the *trailing stop* criterion seems to be better than those of the other two criteria in a sense that the criterion yields larger profits with less losses.

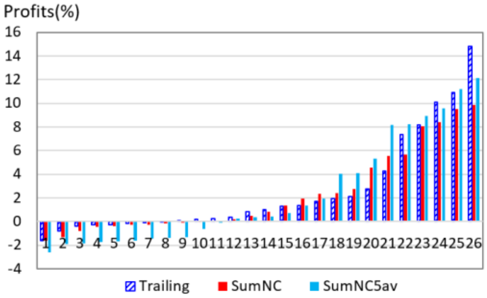


Figure 23. Bar graph of profit and loss for Dow Jones market

Figure 24 shows profit-and-loss patterns in a bar graph for NASDAQ market. Like Dow Jones market, approximately ten out of 26 trade days result in failure. The *SumNC5av* criterion outperforms the other criteria in profits and losses. The maximum profit is about 12%, and the maximum loss is about –3%. Figures 23 and 24 indicate that NASDAQ market is roughly comparable in profit-and-loss patterns to Dow Jones market.

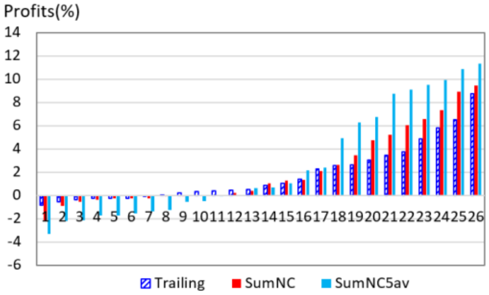


Figure 24. Bar graph of profit and loss for NASDAQ market

Figure 25 shows a profit-and-loss bar graph for Nikkei 225 market. While ten out of 26 trade days are failure like Dow Jones and NASDAQ markets, the maximum profit is about 8%, and the maximum loss is about –4%. The bar graph suggests that Nikkei 225 market seems to have similar price movement patterns, but it is less profitable than Dow Jones and NASDAQ markets.

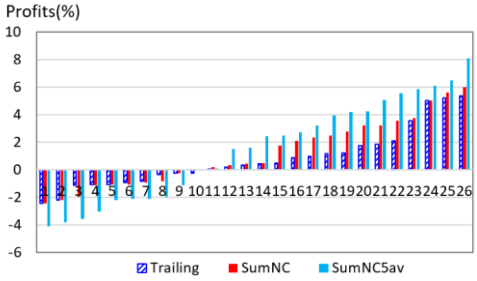


Figure 25. Bar graph of profit and loss for Nikkei 225 market

Figure 26 shows profit-and-loss patterns for Hang Seng market. Roughly, there are 12 out of 26 chances of failure. The maximum profit is estimated about 9%, which is roughly the same as the maximum profit of Nikkei 225 market.

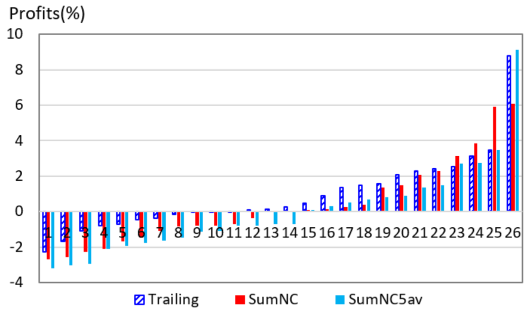


Figure 26. Bar graph of profit and loss for Hang Seng market

Figure 27 shows profit-and-loss patterns for Shanghai market. The results of simulation include a trade on Jan. 4, 2019 with 21.96% profit for the *trailing stop* and *SumND5av* criteria. The trade is deemed to be treated as a special case.

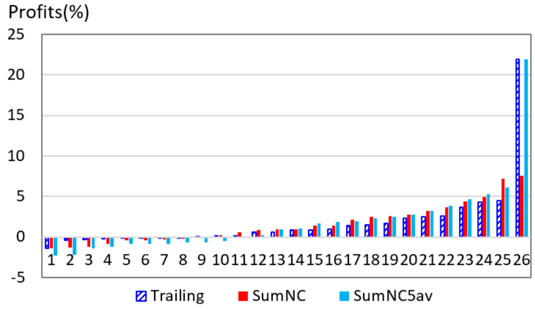


Figure 27. Bar graph of profit and loss for Shanghai market

A feature that is unique in Shanghai market is that profits simulated by the *SumNC* and *SumNC5av* criteria are the same in many cases. The difference between stock price and 5-day average is negative in the Shanghai index as shown in Table II. It is considered that the negative difference between a stock price and a 5-day average leads to the same values of *SumNC* and *SumNC5av* criteria.

I. Summarizing Profit for Each Manket

Table V summarizes "*average profit*", "*success ratio"*, and "*potential profit"* for each criterion. The *trailing stop*, *SumNC*, and *SumNC5av* criteria yield 2.53%, 2.25%, and 2.32% of average profits, respectively, for Dow Jones market as an example. A *success ratio* is obtained by dividing the number of profitable days by the total number of days, i.e., 26. A *potential profit* is calculated by multiplying the *average profit* by the *success ratio*.

TABLE V. SUMMARY OF PROFIT AND SUCCESS RATIO FOR EACH MARKET

	Average Profit(%)			Success Ratio(%)			Potential Profit(%)		
	Trailing	SumNC	SumNC5av	Trailing	SumNC	SumNC5av	Trailing	SumNC	SumNC5av
Dow Jones	2.53	2.25	2.32	69.2	61.5	57.7	1.75	1.38	1.34
NASDAQ	1.78	2.15	2.61	73.1	65.4	57.7	1.30	1.40	1.51
Nikkei 225	0.77	1.21	1.52	61.5	65.4	61.5	0.47	0.79	0.93
Hang Seng	0.90	0.37	0.03	57.7	46.2	46.2	0.52	0.17	0.02
Shanghai	1.84	1.58	1.88	69.2	65.4	61.5	1.27	1.03	1.16

The cell with the largest value among the three criteria is highlighted. The *trailing stop* criterion marks the notable *potential profit* in Dow Jones, Hang Seng, and Shanghai markets. The *SumNC5av* criterion achieves the preferable *potential profit* for NASDAQ and Nikkei 225 markets.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a six-parameter model for retrieving similar candlesticks. The model also deals with the 5-day and 25-day moving averages to identify price trends in addition to a price zone where a stock price occurs. The proposed model is devised to find a buy opportunity. Since a successful stock trade is significantly depends on a sell opportunity, three criteria are proposed and profits that each criterion generates are simulated.

Bullish (uptrend) reversal candlestick patterns consisting of three candlesticks are focused to find a buy opportunity in the empirical study. Three inequality formulas are defined to abstract the bullish reversal patterns. The parameter values used in experiments are determined statistically in terms of the standard deviations of the proposed six parameters to minimize the differences of characteristics among stock markets.

The empirical results show that the US markets, i.e., Dow Jones and NASDAC, are more profitable than Asian markets, i.e., Nikkei 225, Hang Seng, and Shanghai. As for profitability of international markets, the results generally support what is stated in the paper of Dimson, Marsh, and Staunton [3]. The popular *trailing stop* criterion [2] gives the best result among the three exit criteria.

This study only focuses on the bullish reversal patterns in a downtrend, which leads to limitations of the study. Future work may include experiments on the proposed method to measure profitability of other candlestick patterns including bearish (downtrend) reversal patterns that are profitable in short position, and continuation patterns that predict a price trend is likely to remain. Additional studies need to be carried out to measure profitability of global markets to meet demands of finding the most profitable market in the world.

REFERENCES

[1] Y. Udagawa, "Statistical Analysis of Stock Profits to Evaluate Performance of Markets," The Sixth International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2020) IARIA, Feb. 2020, pp. 14–21, ISSN: 2308-4138, ISBN: 978-1-61208-250-9.

[2] C. Mitchell, "Trailing Stop Definition and Uses," Available from: https://www.investopedia.com/terms/t/trailingstop.asp/ May 22, 2020.

[3] E. Dimson, P. Marsh, and M. Staunton, "Should you invest in emerging markets?" London Business School, Available from: https://www.london.edu/think/emerging-markets/ Apr. 2019.

[4] N. Ahmad, R. R. Ahmed, J. Vveinhardt, and D. Streimikiene, "Empirical Analysis of Stock Returns and Volatility: Evidence from Asian Stock Markets," Technological and Economic Development of Economy, vol. 22, Nov. 2016, pp. 808–829.

[5] "Technical Analysis," Cambridge Univ., pp. 1–179, Available from: http://www.mrao.cam.ac.uk/~mph/Technical_Analysis.pdf Feb. 2011.

[6] J. M. Horton, "Stars, crows, and doji: The use of candlesticks in stock selection," Quarterly Review of Economics and Finance, vol. 49, Nov. 2007, pp. 283–294.

[7] R. B. Marshall, R. M. Young, and R. Cahan, "Are candlestick technical trading strategies profitable in the Japanese equity market?" Review of Quantitative Finance and Accounting, vol. 31, Aug. 2008, pp. 191–207.

[8] P. Tharavanij, V. Siraprapasiri, and K. Rajchamaha, "Profitability of Candlestick Charting Patterns in the Stock Exchange of Thailand," SAGE journals, Oct. 2017, pp. 1–18.

[9] G. Caginalp and H. Laurent, "The predictive power of price patterns," Applied Mathematical Finance, vol. 5, Jun. 1998, pp. 181–206.

[10] Y.-J. Goo, D.-H. Chen, and Y.-W. Chang, "The application of Japanese candlestick trading strategies in Taiwan," Investment Management and Financial Innovations, vol. 4, Jan. 2007, pp. 49–79.

[11] C. Chootong and O. Sornil, "Trading Signal Generation Using a Combination of Chart Patterns and Indicators," International Journal of Computer Science Issues, vol. 9, Nov. 2012, pp. 202–209.

[12] C.-F. Tsai and Z.-Y. Quan, "Stock Prediction by Searching for Similarities in Candlestick Charts," Journal ACM Transactions on Management Information Systems (TMIS), vol. 5, Jul. 2014, pp. 1–21.

[13] M. Zhu, S. Atri, and E. Yegen, "Are candlestick trading strategies effective in certain stocks with distinct features?" Pacific Basin Finance Journal, vol. 37, Apr. 2016, pp. 116–127.

[14] M. Jamaloodeen, A. Heinz, and L. Pollacia, "A Statistical Analysis of the Predictive Power of Japanese Candlesticks," Journal of International & Interdisciplinary Business Research, vol. 5, pp. 62–94, Available from: https://scholars.fhsu.edu/jiibr/vol5/iss1/5/ Jun. 2018,

[15] Y. Udagawa, "Dynamic Programming Approach to Retrieving Similar Candlestick Charts for Short-Term Stock Price Prediction," International Journal on Advances in Software, IARIA, vol. 11, Dec. 2018, pp. 440-451.

[16] "Major World Market Indices," Available from: https://www.investing.com/indices/major-indices/ Dec. 2020.

[17] M. Hall, "Introduction to Swing Trading," Available from: https://www.investopedia.com/trading/introduction-to-swing-trading/ Jun. 10, 2020.

[18] S. Glen, "Statistics How To: T-Distribution," StatisticsHowTo.com, Elementary Statistics for the rest of us! Available from: https://www.statisticshowto.com/probability-and-statistics/t-distribution/ 2020.

Effects of UV Irradiation on the Sensing Properties of Co-doped SnO_2 Thin Film for Ethanol Detection

Mikayel Aleksanyan
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: maleksanyan@ysu.am

Vladimir Aroutiounian
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: kisahar@ysu.am

Artak Sayunts
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: sayuntsartak@ysu.am

Valeri Arakelyan
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: avaleri@ysu.am

Hayk Zakaryan
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: hayk.zakaryan@ysu.am

Gohar Shahnazaryan
Department of Physics of Semiconductors and Microelectronics
Yerevan State University
Yerevan, Republic of Armenia
e-mail: sgohar@ysu.am

Abstract - **In this paper, a sputtering ceramic target based on SnO_2 doped with 2 at.% Co was synthesized by solid-phase reaction method. A chemiresistive alcohol vapor sensor based on SnO_2<Co> was manufactured by the high-frequency magnetron sputtering method. The alcohol sensing properties of the SnO_2<Co> sensor under the ultraviolet (UV) illumination were examined at room temperature (RT). The UV-assisted alcohol sensor showed a sufficient response to low concentrations of alcohol vapor at RT. The Co-doped SnO_2 sensor has also demonstrated a high sensitivity to alcohol vapors at elevated operating temperature. The impedance characteristics of the sensors have been also thoroughly studied. It is expected that in the future, Co doped SnO_2 based sensitive thin films will be able to be utilized in highly sensitive, real-time alcohol vapor sensors.**

Keywords - gas sensor; alcohol; UV radiation; room temperature; metal oxides; Nyquist plot.

I. INTRODUCTION

Today, alcohol vapor sensors have a great demand in various fields. Ethanol sensors are used in the food industry, medicine and biotechnology. Ethanol sensors are also extremely important during the production of ethanol and alcoholic drinks to monitor the beverage quality. They are used in processes such as: food–packaging, clinical analysis, agronomic, vinicultural and veterinary analysis, also toxic waste and contamination analysis, fuel processing, Trends in Analytical Chemistry (TRAC) management and societal applications, as well as chemical processing in industry [1]-[6]. Several methods and strategies have been reported for the detection of ethanol, e.g., gas chromatography, liquid chromatography, refractometry and spectrophotometry, semiconductor gas sensors and so on [3] [7] [8].

The solid-state gas sensors based on Metal Oxide Semiconductors (MOSs) with different nanostructures have played an important role in environmental monitoring, domestic and car safety, control in chemical processing due to their distinct advantages, such as simple implementation, low cost, high sensitivity, stability and reproducibility, low detection limit, easy production, nontoxicity, easy-achieved real-time response and compatibility with micro-fabrication processes [9]-[11]. Various MOSs materials, such as SnO_2, In_2O_3, WO_3, ZnO, TiO_2, Fe_2O_3, CuO, Ga_2O_3, CTO (CrTiO) with different nanostructures and dopant have been studied and showed promising results for detecting Volatile Organic Compounds (VOCs) [12]. Among these materials, the SnO_2 has good electrical and chemical properties. It is an n-type semiconductor with tetragonal rutile structure and it has a large energy band gap of 3.6 eV at 300 K. It has been widely exploited as an ultrasensitive gas sensor for the detection of carbon monoxide (CO), ammonia, ozone, carbon dioxide, hydrogen, hydrogen peroxide, nitrogen dioxide, ethanol and so on [13]-[15]. The wide range of possible applications has attracted many researchers to work on this material with different nanostructures, such as nanograins, nanorods, nanowires and nanofibers synthesized by various methods. It has a high sensitivity to reducing and oxidizing gases, fast response and recovery behavior and low sensitivity to humidity [16]-[18].

Although many conductometric gas sensors made of MOSs have been commercialized for the last decades, a lot of problems still need to be solved in order to improve the performance of gas sensing devices. The main issues are related to sensitivity, selectivity and stability but the lowering of sensor's operating temperature is still one of the main concerns. Resistive metal oxide based gas sensors normally operate at an elevated temperature (in a range of 200 ºC to 400 ºC). This results in higher power consumption, limits the use of the sensor in explosive environments, and causes difficulties for the sensor to be attached to electrical systems [19] [20].

There are many studies aimed at applying new technologies and reducing operating temperatures. To ensure a low operating temperature, several techniques have been used, such as doping the metal oxides with additives, using catalytic particles, applying a high electric field across the sensor terminals and illuminating the sensors with UV radiation [21] [22]. The irradiation of UV-assisted MOS sensors is an important alternative to activate chemical reactions on the metal oxide surface and reduce the resistance of the thin sensing layer instead of the more common use of energy-demanding heating. Almost completely replacing the effect of thermal energy, UV irradiation greatly influences the adsorption and desorption processes of the gas on the semiconductor surface enhancing their reactivity with the analyte gas. Under the influence of UV illumination, as a result of the formation of electron–hole pairs, more neutral atoms and molecules of absorbed oxygen on the surface of the semiconductor become ions, which then interact with analyte gas. UV irradiation can also be used to clean the active surface of a gas sensing layer, but the more important function is to improve the sensitivity and selectivity of the gas sensor by reducing the operating temperature. If it is not possible to lower the operating temperature to RT by using UV irradiation, UV irradiation combined with heating can be used to stimulate the gas sensor [23]-[25].

In this paper, we focus on low temperature sensing of SnO_2 based thin film sensors under UV illumination. In Section II, the fabrication steps of SnO_2<Co> sensor are presented. In Section III, the studies of sensing properties of UV assisted ethanol sensor are presented. In Section IV, the gas sensing mechanisms are explained. The conclusions are outlined in Section V. The sensor exhibited good sensitivity to low concentration of ethanol vapors. Fabricated sensors have also sufficient selectivity and stability over time.

II. Sensor fabrication

Sensitive layers based on SnO_2<Co> were deposited by the RF magnetron sputtering technique. Firstly, appropriate quantities of the corresponding metal oxide powders (SnO_2+2 at.%Co_2O_3) were weighed and mixed thoroughly for 10 hours. Then, the mixture was subjected to pre-heat treatment at 800 ^{0}C for 5 hours (the initial annealing temperature was chosen based on the composition of the compound). The preheating of mixed powder eliminates the moisture of the metal oxide raw materials, which facilitates homogeneous mixing and milling of the powders (when the ceramic tablet is made of dry powders, it reduces the probability of the formation of mechanical cracks during final annealing). Then, the mixed powder was milled for 20 hours until becoming fully homogeneous and pressed (with 2000 N/cm^2 pressure) in a form of a tablet (with 50 mm diameter). The sputtering ceramic target based on SnO_2 doped with 2 at.% Co (using the pressed tablet) was synthesized by solid-phase reaction method using thermal treatment in the atmosphere by the programmable furnace Nabertherm, HT O4/16 (with the controller of C 42). The final annealing was carried out at temperature range of 500 °C-1100 °C for 20 hours. The synthesized semiconductor solid solution was subjected to mechanical treatment in order to eliminate surface defects. So, a smooth and parallel target with a diameter of 40 mm and thickness of 2 mm was prepared as a magnetron sputtering target (see Figure 1).

The thin sensing layers were deposited on Multi-Sensor-Platforms by the RF magnetron sputtering method using synthesized SnO_2<Co> target. The Multi-Sensor-Platforms were purchased from TESLA BLANTA (Czech Republic). The platform has a temperature sensor (Pt 1000) for controlling the operating temperature. There are platinum heater and interdigitated electrodes on the ceramic substrate of the Multi-Sensor-Platform (see Figure 2). The heater and temperature sensor were covered with an insulating glass layer. Gas sensitive SnO_2<Co> layer was deposited onto the non-passivated electrode structure, so the Multi-Sensor-Platform was converted into a gas sensor. Then, palladium catalytic particles are deposited on the surface of the magnetron sputtered sensing layer the by ion-beam sputtering method for sensitization of the active layer. The working conditions of the high-frequency magnetron sputtering and ion-beam sputtering are presented in Table I (the base pressure was 2×10^4 Pa for both cases). The manufactured sensors were annealed in the air at 350 °C for 4 hours for homogenization of sensing films and stabilizing their parameters. The fabrication steps of photo-assisted gas sensors are presented in Figure 1.

The thickness of the SnO_2<Co> thin film was measured by the Alpha-Step D-300 (KLA Tencor) profiler. The result of the study of the film-substrate transition profile is shown in Figure 3. The thickness of the SnO_2<Co> film was equal to 180 nm.

The electrical and gas sensing properties of the SnO_2<Co> thin layer was measured using a home-made computer-controlled gas testing system. The testing system has a test chamber, pressure sensor (Motorola-MPX5010DP) and a data acquisition system (PCLD-8115) [26]. For measurement of alcohol vapor concentration, the SnO_2<Co> based sensor (the Multi-Sensor-Platform) was attached in the test chamber connecting the six pins (two pins of temperature sensor, two pins of heater and two pins of resistance measurement electrodes, see Figure 2) with the corresponding inputs on sensor holder. The UV LED (λ=365 nm) was attached 0.5 cm away from the active layer with an illumination of 2 mW/cm^2. The gas sensing

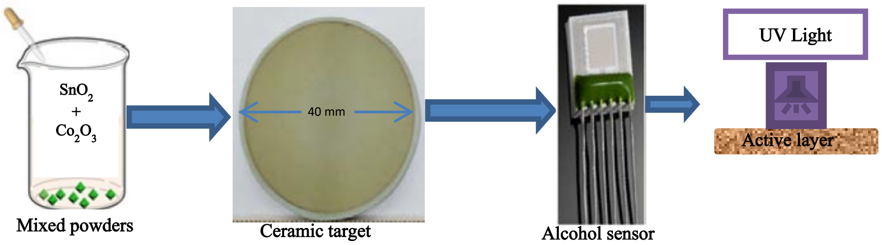


Figure 1. Schematic block diagram of the photo-assisted gas sensor fabrication.

TABLE I. THE WORKING CONDITIONS FOR DEPOSITION OF THIN LAYER AND CATALITIC PARTICLES.

Process	Sputtering duration	Working pressure	Power of generator	Substrate temperature	Cathode current	Anode voltage	Sputtering gas
Magnetron sputtering (RF)	20 m	2×10^{-1} Pa	60 w	200 ^{0}C	---	---	Ar
Ion-beam sputtering (DC)	3 s	5×10^{-1} Pa	---	100 ^{0}C	65 A	25 V	Ar

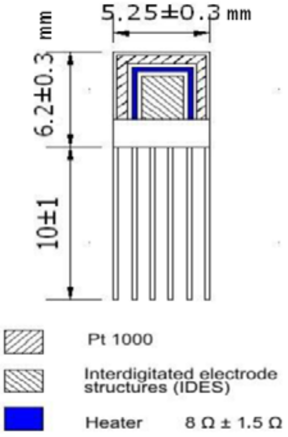


Figure 2. The schematic diagram of the Multi-Sensor-Platform.

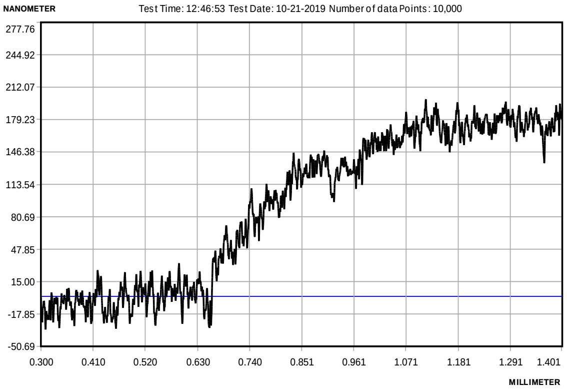


Figure 3. The thickness measurement result for the Co-doped SnO_2 film.

properties of the SnO_2<Co> sensor was measured at RT in the dark and under UV illumination. The response of the sensor was also measured at 200 ^{0}C operating temperature in the dark. The working temperature of the sensor was adjusted by changing the voltage across the platinum heater.

To have the necessary concentration of alcohol vapor in the chamber, the liquid ethanol was introduced into the chamber on the special hot plate designed for the quick conversion of the liquid ethanol to the gas phase. The response of the sensor is defined as $[(R_a-R_g)/R_a]\times100$ %, where R_a and R_g are the electrical resistances of active layer in air and target gas, respectively.

III. GAS SENSING PERFORMANCES

Initially, we tested the influence of the UV illumination on the baseline resistance of the SnO_2<Co> sensor at RT. It can be seen from Figure 4 that the value of R_0/R_{UV} (~350) ratio is larger than 1, indicating the decrease of the sensor baseline resistance under UV illumination.

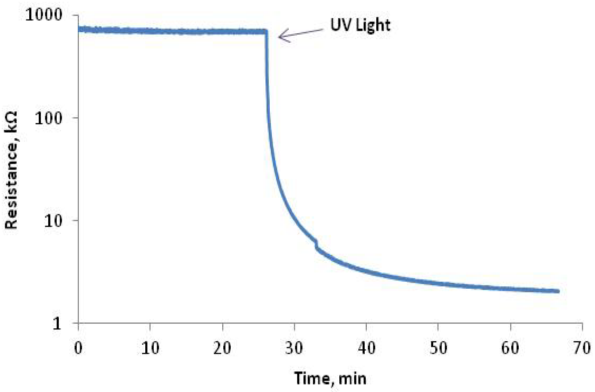


Figure 4. Resistance variation of the Co-doped SnO_2 sensing layer under the influence of UV irradiation at RT.

UV rays generate free carrier in the semiconductor, as a result of which the baseline resistance decreases (These processes are discussed in more detail in Section III). The response time of the Co-doped SnO_2 thin film under UV irradiation is a few minutes.

The manufactured sensor is resistive and its operation is grounded on changes of resistance of gas sensitive semiconductor layer under the influence of ethanol vapors caused by an exchange of charges between molecules of the semiconductor film and absorbed ethanol. The high operating temperature of these types of sensors is mainly due to the high activation energies of chemical reactions. For this reason, these types of sensors are mainly not sensitive at RT. The UV light promotes the gas adsorption and desorption on the surface of the semiconductor participating to the sensing mechanisms [24].

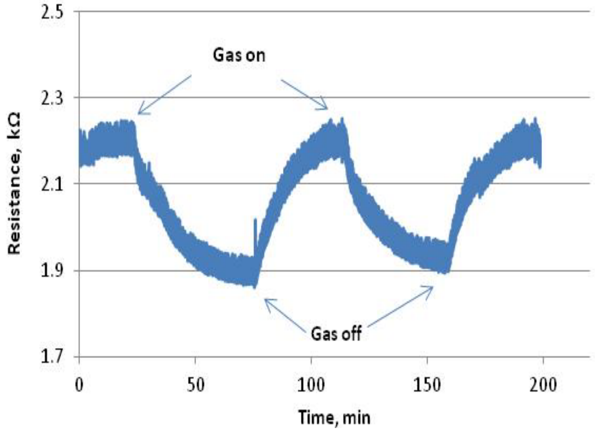


Figure 5. Resistance variation of the SnO_2<Co> sensor under the influence of UV irradiation at RT in the presence of 150 ppm ethanol vapors.

The thin film SnO_2<Co> based sensor did not show sensitivity to ethanol vapors at RT without UV irradiation. We measured the resistance variation (also the signal repeatability) of the SnO_2<Co> sensor in the presence of ethanol vapors under the influence of UV irradiation at RT. The resistance of the thin film changes by almost 400 Ω in the presence of 150 ppm ethanol vapors (see Figure 5).

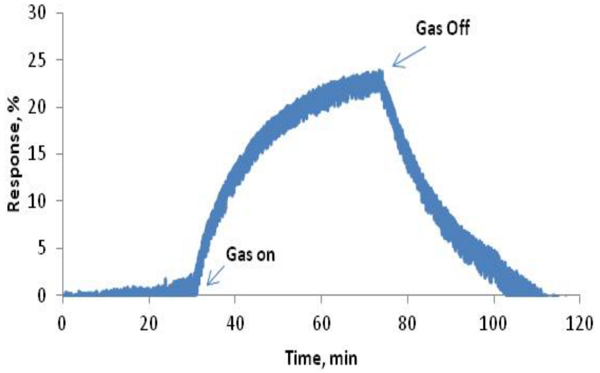


Figure 6. The SnO_2<Co> sensor's response to 900 ppm of ethanol vapors under the influence of UV irradiation at RT.

Sensor response and recovery times are in minutes and it is clear that recovery times are faster because UV light more stimulate the desorption processes from the surface of the sensing layer.

Figure 6 shows the transient response of the SnO_2<Co> sensor in the presence of ethanol vapors under UV light at RT. The response to 900 ppm ethanol vapors under UV illumination is sufficiently high (24 %).

We extracted the response vs. concentration curve for the Co-doped SnO_2 sensitive film. Figure 7 shows the dependence of response on the ethanol vapor concentration under the influence of UV irradiation at RT. The dependence has almost linear characteristic, which will allow not only to detect of ethanol vapors but also to accurately measure the low concentrations of this gas.

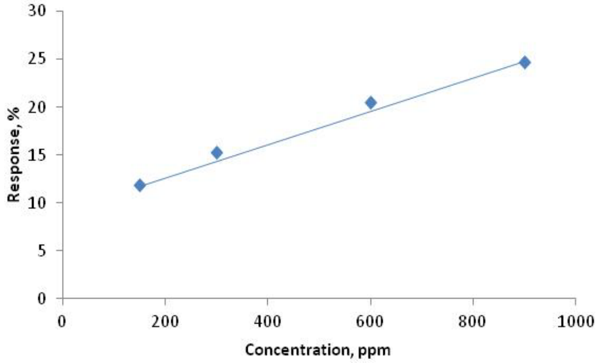


Figure 7. The dependence of response on the ethanol vapor concentration under the influence of UV irradiation at RT.

The gas sensing properties of the Co-doped SnO_2 sensor under the influence of ethanol vapors in dark conditions at high operating temperatures and under the influence of UV irradiation combined with heating were also studied. The sensor's responses to ethanol vapors at different operating temperatures were initially measured with UV irradiation.

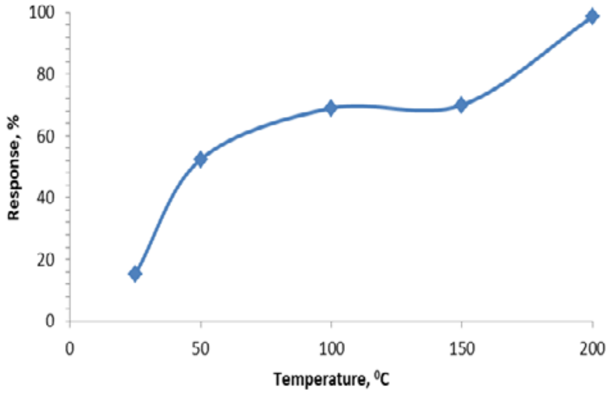


Figure 8. The dependence of response on the operating temperature of the SnO_2<Co> sensor under the influence of UV irradiation at the presence of 300 ppm ethanol vapors.

Figure 8 shows the dependence of response on the operating temperature of the SnO_2<Co> sensor under the influence of UV irradiation at the presence of 300 ppm ethanol vapors. As expected, in the case of not very high operating temperatures (up to 200 ^{0}C), the increase in temperature is accompanied by an increase in the response, because the chemical reactions at higher operating temperatures are faster and more efficient.

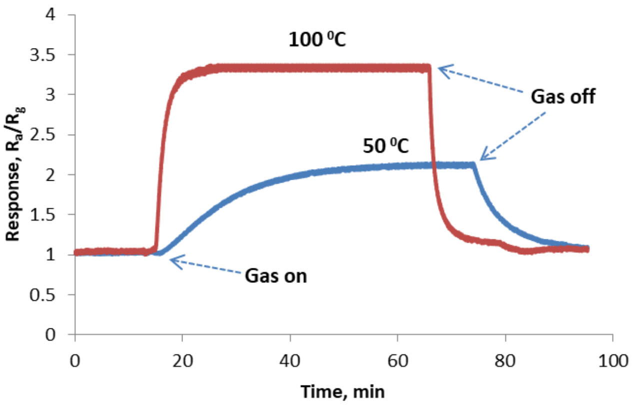


Figure 9. The SnO_2<Co> sensor's responses to 300 ppm of ethanol vapors under the influence of UV irradiation at different operating temperature.

The sensor's responses were also compared at different operating temperatures under the influence of UV irradiation in dark conditions. It should be noted that the sensor did not show sensitivity to alcohol vapors at 50 ^{0}C and 100 ^{0}C operating temperatures in dark conditions. The SnO_2<Co> sensor's responses to 300 ppm of ethanol vapors under the influence of UV irradiation at 50 ^{0}C and 100 ^{0}C operating temperatures are presented in Figure 9 (At high operating temperatures, as we are dealing with higher response values, it is desirable to use an absolute response definition: R_a/R_g). The effect of the UV irradiation at these temperatures gave the sensor increased sensitivity and as expected, at 100 ^{0}C we had a higher response and shorter recovery and response times.

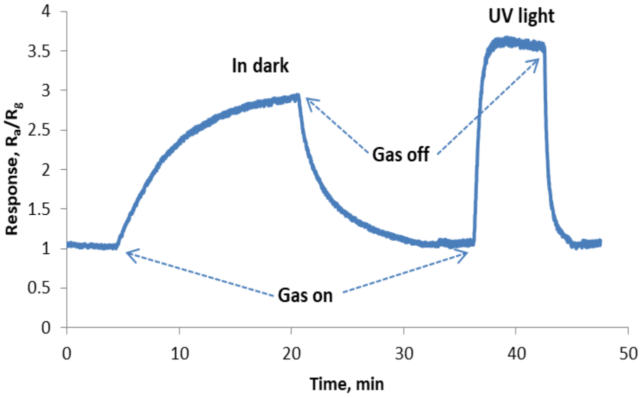


Figure 10. The SnO_2<Co> sensor's responses to 300 ppm of ethanol vapors at 150 ^{0}C operating temperature in dark conditions and under the influence of UV irradiation.

The SnO_2<Co> sensor's responses to 300 ppm of ethanol vapors at 150 ^{0}C operating temperature in dark conditions and under the influence of UV irradiation are presented in Figure 10. The effect of the UV irradiation dramatically improves the speed and the response of the sensor.

Since the maximum response in the observed operating temperature range (25-200 ^{0}C) was recorded at 200 ^{0}C operating temperature, the gas sensitivity characteristics of the sensor were studied in more detail at this temperature. The sensitive layer's resistance decreases more than 25 times in the presence of 150 ppm ethanol vapors at 200 ^{0}C operating temperature (see Figure 11). The response and recovery times of the sensor at high operating temperatures are a few seconds.

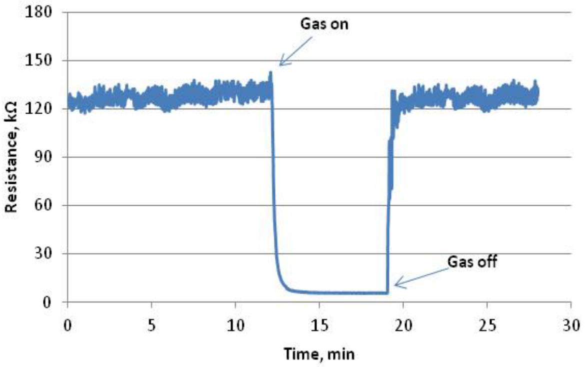


Figure 11. Resistance variation of the SnO_2<Co> sensor in the presence of 150 ppm ethanol vapors at 200 °C operating temperature in dark conditions.

The sensor showed sensitivity (R_a/R_g=3) to the extremely low concentrations (0.5 ppm) of ethanol vapors at 200 ^{0}C operating temperature even in dark conditions. The presence of the UV irradiation increases the response to 3.5 and reduces the recovery time (see Figure 12).

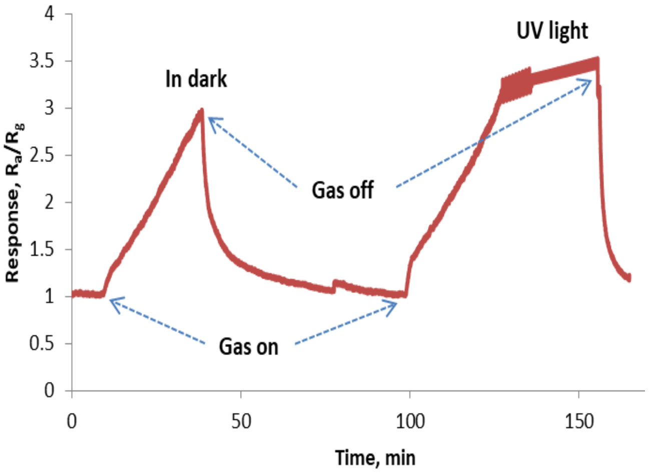


Figure 12. The SnO_2<Co> sensor's responses to 0.5 ppm of ethanol vapors at 200 ^{0}C operating temperature in dark condition and under the influence of UV irradiation.

Figure 13 shows the SnO_2<Co> sensor's responses to different concentrations of ethanol vapors at 200 ^{0}C operating temperature under the influence of UV irradiation. The sensor's response to 300 ppm ethanol vapors is 75, which is extremely high. The response and recovery times of the sensor at 200 ^{0}C operating temperature are a few seconds at relatively high concentrations (300 ppm and 100 ppm) but, as the concentration decreases, the response times increase. It is assumed that, in the case of low concentrations, the saturation of the gas-sensitive surface with ethanol molecules occurs delayed, which leads to an increase in the response time of the sensor.

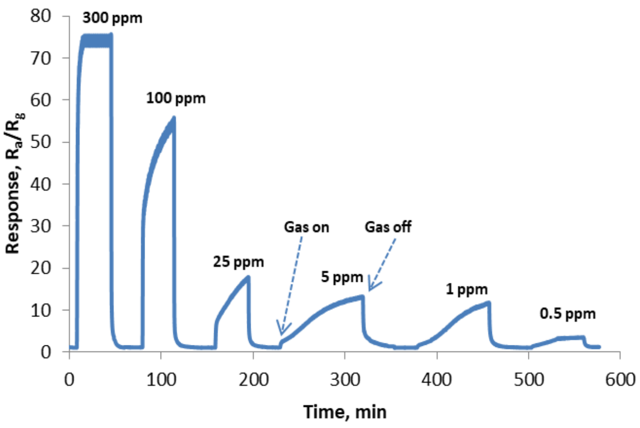


Figure 13. The SnO$_2$<Co> sensor's responses to different concentrations of ethanol vapor at 200 ^{0}C operating temperature under the influence of UV irradiation.

The response vs. concentration curve of the sensor was also extracted under the influence of UV irradiation at 200 ^{0}C operating temperature (Figure 14). The dependence has almost linear characteristic at the concentration range of 0.5 to 100 ppm, but in the case of higher concentration, we have a deviation from the initial linear curve. At higher concentration, the angle of the linear curve changes and it is expected, that it will also be linear (which will show our future experimental work).

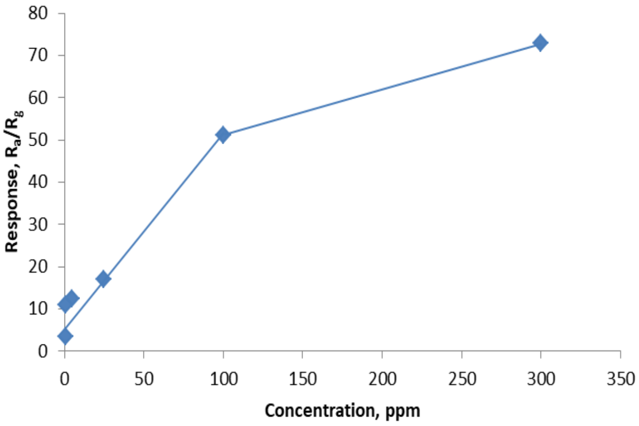


Figure 14. The dependence of response on the ethanol vapor concentration under the influence of UV irradiation at 200 °C operating temperature.

So, if it is not possible to lower the operating temperature to RT by using UV irradiation, UV irradiation combined with heating can be used to stimulate the gas sensor. At high operating temperature both in dark conditions and with UV irradiation, the sensor performance is quite promising but the power consumption of fabricated sensor at 200 ^{0}C is about 2.5 W. It is more than two orders high then the power consumption (24 mW) needed the sensor operating with UV irradiation at RT.

The alcohol responses of the SnO$_2$<Co> based sensor were also compared with those described in previous reports (Table II). Our UV-activated SnO$_2$<Co> based sensor exhibits much lower working temperature and comparable ethanol response compared with the previously reported ethanol sensors with and without UV illumination. Our UV activated SnO$_2$<Co> sensor displays better ethanol response to extremely low concentration (0.5 ppm) of ethanol than most of the reported oxide-based sensors under UV irradiation.

TABLE II. THE COMPARISON OF ETHANOL VAPOR SENSOR RESPONSE BETWEEN THIS WORK AND PREVIOUSLY PUBLISHED REPORTS.

Sensing materials	Conc. (ppm)	Temp. (^{0}C)	Resp.	UV light	Ref.
SnO$_2$-Zn$_2$SnO$_4$	200	300	5	No	[27]
ZnO	150	53	1.7	Yes	[28]
Zn$_2$SnO$_4$	200	130	32.5	Yes	[29]
ZnO:AuNPs	1000	125	6.3	Yes	[30]
SnO$_2$-ZnO	100	160	1.1	No	[31]
SnO$_2$-GaN	500	RT	1.01	Yes	[32]
NiO	500	200	4.94	Yes	[33]
SnO$_2$	300	240	65	Yes	[34]
SnO$_2$<Co>	900	RT	1.4	Yes	This work
SnO$_2$<Co>	0.5	200	5	Yes	This work

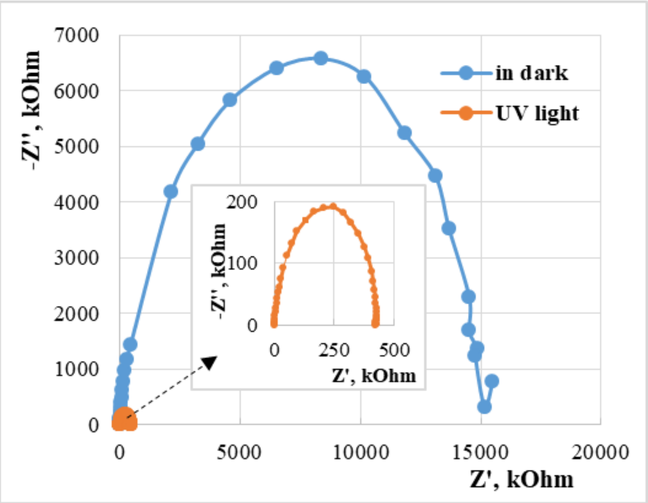


Figure 15. The Nyquist plots of SnO$_2$<Co> sensor observed under dark condition and in the presence of UV irradiation at RT.

There were also studied the gas sensing properties of the Co-doped SnO$_2$ sensor by impedance spectroscopy using the ZIVE-SP1 Potentiostat and the Keithley 4200-SCS (Semiconductor Characterization System) under the influence of ethanol vapors in dark conditions at high operating temperatures and under the influence of UV irradiation combined with heating.

Figure 15 shows the Nyquist plots of SnO_2<Co> sensor observed under dark condition and in the presence of UV irradiation at RT. The UV response is quite large presented by the deviation of the semicircular Nyquist plots.

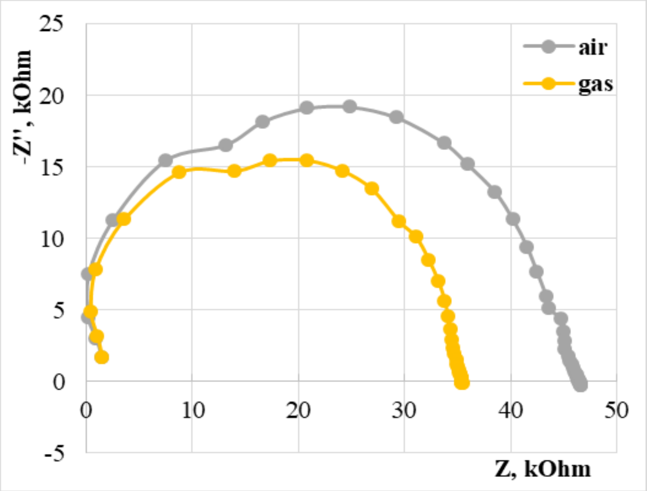


Figure 16. The Nyquist plots of SnO_2<Co> sensor observed in the air and in the presence of 300 ppm ethanol vapors at 50 ^{0}C operating temperature under the influence of UV irradiation.

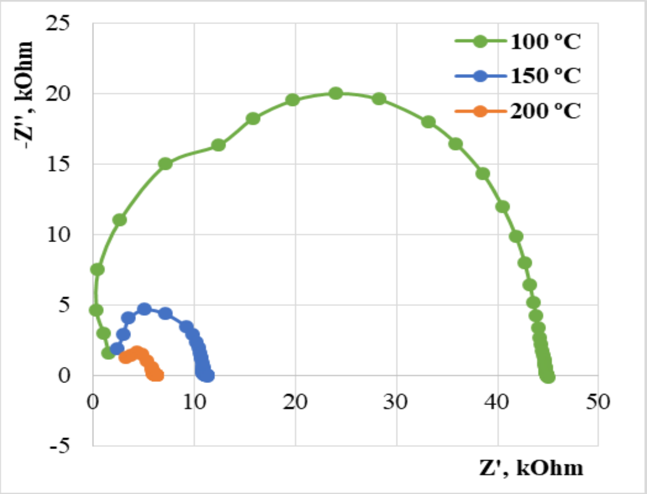


Figure 17. The Nyquist plots of SnO_2<Co> sensor observed in the presence of 300 ppm ethanol vapors at different operating temperatures without UV irradiation.

The sensor did not show sensitivity to alcohol vapors at 50 ^{0}C in dark conditions, but at this operating temperature under the influence of UV irradiation the response was significant. The Nyquist plots of the sensor observed in the air and in the presence of 300 ppm ethanol vapors at 50 ^{0}C operating temperature under the influence of UV irradiation are presented in Figure 16. The significant deviation of Nyquist plots at the presence of ethanol vapors was observed.

The Nyquist plots of SnO_2<Co> sensor observed in the presence of 300 ppm ethanol vapors at different operating temperatures under the influence of UV irradiation and in the dark condition are presented in Figure 17 and Figure 18. The semicircle plots were obtained for Nyquist impedance, which indicates that the diameter of the semicircles increased gradually when the temperature was decreased. It is assumed that the deviation of the semicircles from the zero point by the influence of UV irradiation is related to the direct effect of the UV-activated adsorption/desorption processes on the impedance properties.

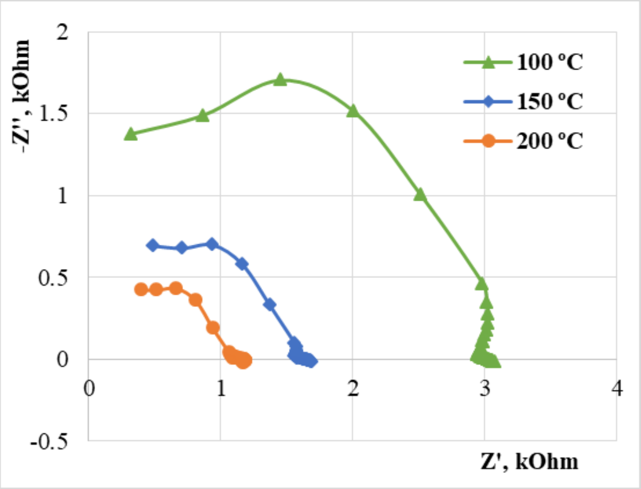


Figure 18. The Nyquist plots of SnO_2<Co> sensor observed in the presence of 300 ppm ethanol vapors at different operating temperatures under the influence of UV irradiation.

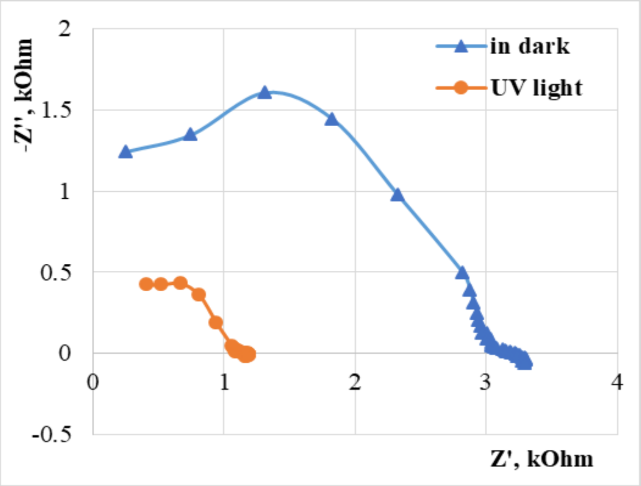


Figure 19. The Nyquist plots of SnO_2<Co> sensor observed in the presence of 300 ppm ethanol vapors at 200 ^{0}C operating temperature under dark condition and in the presence of UV irradiation.

Figure 19 shows that Nyquist impedance of SnO_2<Co> sensor recorded in the presence of 300 ppm ethanol vapors at 200 ^{0}C operating temperature under dark condition and in the presence of UV irradiation. The impact of the UV irradiation dramatically reduced the diameter of the semicircle by shifting the curve to lower ranges of the resistance. This is due to the change of the localized charges and free carriers concentration on the active surface of the semiconductor.

The frequency dependencies for the real and imaginary components of the impedance for the SnO_2<Co> sensor were also extracted (see Figure 20 and Figure 21). The deflection of the curves under the influence of UV light in the case of the imaginary components of the impedance depends more on the frequency. It is clear from the Figure 21 that the deviation is more significant in the case of high frequencies. This is due to the fact that at high frequencies the capacitive properties of the sensitive film are revealed.

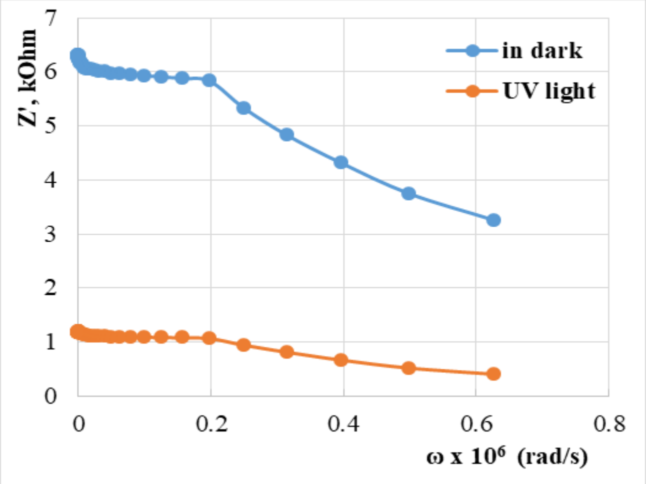


Figure 20. Frequency dependencies for the real component of the impedance for the SnO_2<Co> in the presence of 300 ppm ethanol vapors at 200 ^{0}C operating temperature under dark condition and in the presence of UV irradiation.

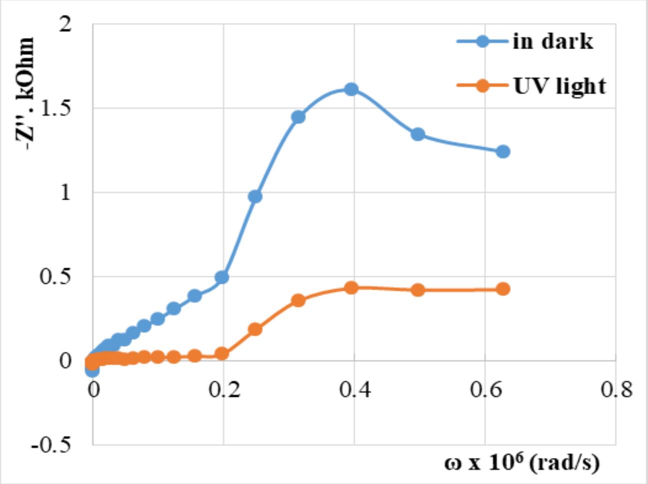


Figure 21. Frequency dependencies for the imaginary component of the impedance for the SnO_2<Co> in the presence of 300 ppm ethanol vapors at 200 ^{0}C operating temperature under dark condition and in the presence of UV irradiation.

Our future researches will focus not only on the study of the deviations of Nyquist plots of the SnO_2<Co> sensor at the presence of ethanol vapors under the influence of UV irradiation, but also on the build of the sensor equivalent circuit, obtaining more comprehensive information about the gas-sensitive element.

The sensor's responses to acetone and toluene vapors were also measured. The sensor showed negligible sensitivity to acetone vapors, but did not show any sensitivity to toluene vapors, thus, the sensor has a high selectivity in the presence of VOCs.

The fabricated sensors can be also easily attached to modern Arduino systems for data extraction and evaluation.

IV. Gas Sensing Mechanism

To explain the observed sensing behavior of the Co-doped SnO_2 sensor, the gas sensing mechanism under dark and UV light conditions has to be taken into account. The basis of the operation of conductometric sensors is the change in resistance under the effect of reactions taking place on the surface of the sensing layer [35]-[37]. The target gases (chemical species) interact with the sensitive layer and thus modulate its electrical conductance. The gas sensing mechanism includes consideration of the role of the chemisorbed oxygen. The initial exposure to air results in oxygen adsorption on the surface through transferring electrons from the conduction band to the adsorbed oxygen. The oxygen chemisorption means the formation of O^{2-}, O^- and O_2^- species on the surface. They originate due to electrons which are captured by adsorbed neutral oxygen species on the surface of the oxide. For the n-type semiconductor the majority charge carriers are electrons and upon interaction with a reducing gas an increase in conductivity occurs [38]-[40]. The oxygen ions are adsorbed mainly molecularly (O_2^-) in the absence of UV radiation and atomic oxygen ions (O^- and O^{2-}) may be formed on the surface of the illuminated sensor, as shown in the following reactions [24] [41]:

$$O_{2\,(gas)} \leftrightarrow O_{2\,(adsorbed)} \tag{1}$$

$$O_{2\,(adsorbed)} + e^- \leftrightarrow O^-_{2\,(adsorbed)} \tag{2}$$

$$O_{2(gas)} + e^-_{(hv)} \leftrightarrow O^-_{2\,(adsorbed)} \tag{3}$$

$$O^-_{2\,(adsorbed)} + e^-_{(hv)} \leftrightarrow 2O^-_{(adsorbed)} \tag{4}$$

$$O^-_{(adsorbed)} + e^-_{(hv)} \leftrightarrow O^{2-}_{(adsorbed)} \tag{5}$$

UV illumination changes the number of charge carriers on the surface of the film through exciting electrons from the material valence band to the conduction band, which results in a decrease in sensor resistance and an increase of the number of surface atomic oxygen ions. The oxygen chemisorption results in a modification of the space charge region toward depletion.

Upon exposure to alcohol vapors, the ethanol molecules react with surface oxygen species and produce electrons, resulting in an increase of electrical conductance of the n-type semiconductor (Co-doped SnO_2 sensitive film). The appropriate reactions are expressed as follows [42]-[44]:

$$6O^{-}_{(adsorbed)} + C_2H_5OH_{(gas)} \rightarrow 3H_2O_{\ (gas)} + 2CO_{2\ (gas)} + 6e^{-} \quad (6)$$

$$3O_2^{\ -}{}_{(adsorbed)} + C_2H_5OH_{(gas)} \rightarrow 3H_2O_{(gas)} + 2CO_{2(gas)} + 3e^{-} \quad (7)$$

$$6O^{2-}{}_{(adsorbed)} + C_2H_5OH_{(gas)} \rightarrow 3H_2O_{(gas)} + 2CO_{2(gas)} + 12e^{-} \quad (8)$$

The continuous UV illumination promotes the formation of more atomic oxygen ions (O^{-} and O^{2-}) on the surface of the semiconductor, which leads to increased sensitivity.

V. Conclusion

In summary, a simple technology has been used to manufacture semiconductor thin film sensor based on SnO_2 doped with 2 at.% Co. The fabricated SnO_2<Co> chemiresistive gas sensor showed a good sensitivity to different concentrations of ethanol vapor (from 150 to 900 ppm) at RT with the activation of low-powered UV LED (24 mW, 365 nm). The sensor also showed sensitivity to extremely low concentrations (0.5 ppm) of ethanol vapors at 200 ^{0}C operating temperature even in dark condition and the presence of the UV irradiation increases the response and reduces the recovery time. The sensor displayed a good signal repeatability and long-term stability. The sensor provides not only high response to ppm level of alcohol vapors but also significant deviation of Nyquist plots at the presence of alcohol vapors. These sensing characteristics made the present SnO_2<Co> based sensor a promising candidate for practically detecting ethanol vapors at the temperature range of 25 to 200 ^{0}C.

Acknowledgment

This investigation was supported by 19YR-2K002 (Young Researchers 2019-2021) project of Ministry of Education, Science, Culture and Sport RA (Science Committee).

References

[1] M. Aleksanyan, A. Sayunts, H. Zakaryan, V. Aroutiounian, V. Arakelyan, and G. Shahnazaryan, "UV-assisted Chemiresistive Alcohol Sensor Based on Cobalt Doped Tin Dioxide," The Fifth International Conference on Advances in Sensors, Actuators, Metering and Sensing (ALLSENSORS 2020) IARIA, Nov. 2020, pp. 1-9, ISBN 978-1-61208-766-5.

[2] M. A. Lakhanea, A. L. Choudharia, R. S. Khairnara, and M. P. Mahabolea, "Alcohol Sensor Based on Mg-STI zeolite Thick Films," Procedia Technology, vol. 24, pp. 595-602, 2016.

[3] A. Charishma, A. Jayarama, V. V. D. Shastrimath, and R. Pinto, "An Ethanol Sensor Review: Materials, Techniques and Performance," SAHYADRI International Journal of Research, vol. l 3, pp. 37-46, June 2017.

[4] E. C. Ramaa et al., "Comparative study of different alcohol sensors based on Screen-Printed Carbon Electrodes," Analytica Chimica Acta, vol. 728, pp. 69-76, 2012, doi:10.1016/j.aca.2012.03.039.

[5] Y. Li et al., "In situ decoration of Zn_2SnO_4 nanoparticles on reduced graphene oxide for high performance ethanol sensor," Ceramics International, vol. 44, pp. 6836–6842, January 2018, doi:org/10.1016/j.ceramint.2018.01.107.

[6] Z. Qin et al., "Highly sensitive alcohol sensor based on a single Er-doped In_2O_3 nanoribbon," Chemical Physics Letters, vol. 646, pp. 12–17, 2016, doi:10.1016/j.cplett.2015.12.054.

[7] G. Feng, M. Zhang, S. Wang, C. Song, and J. Xiao, "Ultra-fast responding and recovering ethanol sensors based on CdS nanospheres doped with graphene," Applied Surface Science, vol. 453, pp. 513–519, May 2018, doi:10.1016/j.apsusc.2018.05.102.

[8] M. Shashikant, V. Lahade1, Mr. Pravin, and D. Pardhi, "Gas Sensing Technologies: Review, Scope and Challenges," International Journal of Recent Trends in Engineering & Research (IJRTER), vol. 04, pp. 108-115, February 2018, doi: 10.23883/IJRTER.2018.4073.M2XNS.

[9] Y. Shen et al., "Highly sensitive and selective room temperature alcohol gas sensors based on TeO_2 nanowires," Journal of Alloys and Compounds, vol. 664, pp. 229-234, 2016, doi: 10.1016/j.jallcom.2015.12.247.

[10] V. M. Aroutiounian et al., "Manufacturing and investigations of i-butane sensor made of SnO_2/multiwall-carbon-nanotube nanocomposite," Sensors and Actuators B, vol. 173, pp. 890-896, 2012, doi:10.1016/j.snb.2012.04.039.

[11] V. M. Arakelyan et al., "Gas sensors made of multiwall carbon nanotubes modified by tin dioxide," Journal of Contemporary Physics (Armenian Academy of Sciences), vol. 48, pp. 176-183, 2013, doi:10.3103/S1068337213040063.

[12] G. Korotcenkov and B. K. Cho, "Metal oxide composites in conductometric gas sensors: Achievements and challenges," Sensors and Actuators B, vol. 244, pp. 182-210, June 2017, doi:10.1016/j.snb.2016.12.117.

[13] V. Aroutiounian et al., "Thin-film SnO_2 and ZnO detectors of hydrogen peroxide vapors," Journal of Sensors and Sensor Systems, vol. 7, pp. 281-288, April 2018, doi: /10.5194/jsss-7-281-2018.

[14] H. S. Jeong, M. J. Park, S. H. Kwon, H. J. Joo, S. H. Song, and H. I. Kwon, "Low temperature NO_2 sensing properties of RF-sputtered SnO-SnO_2 heterojunction thin-film with p-type semiconducting behavior," Ceramics International, vol. 44, pp. 17283–17289, June 2018, doi:10.1016/j.ceramint.2018.06.189.

[15] V. Aroutiounian et al., "Nanostructured sensors for detection of hydrogen peroxide vapours," Sensors & Transducers, vol. 213, pp. 46-53, June 2017.

[16] G. Korotcenkov and V. Nehasil, "The role of Rh dispersion in gas sensing effects observed in SnO_2 thin films," Materials Chemistry and Physics, vol. 232, pp. 160-168, June 2019, doi:10.1016/j.matchemphys.2019.04.069.

[17] G. Korotcenkov and B. K. Cho, "Thin film SnO_2-based gas sensors: Film thickness influence," Sensors and Actuators B, vol. 142, pp. 321-330, October 2009, doi:10.1016/j.snb.2009.08.006.

[18] M. S. Aleksanyan, "Methane sensor based on $SnO_2/In_2O_3/TiO_2$ nanostructure," Journal of Contemporary Physics (Armenian Academy of Sciences), vol. 45, pp. 77-80, 2010, doi:10.3103/S1068337210020052.

[19] N. Li, Y. Fan, Y. Shi, Q. Xiang, X. Wang, and J. Xu, "A low temperature formaldehyde gas sensor based on hierarchical SnO/SnO_2 nano-flowers assembled from ultrathin nanosheets: Synthesis, sensing performance and mechanism," Sensors and Actuators B, vol. 294, pp. 106-115, September 2019, doi:10.1016/j.snb.2019.04.061.

[20] I. Kortidis, H. C. Swart, S. S. Ray, and D. E. Motaung, "Detailed understanding on the relation of various pH and synthesis reaction times towards a prominent low temperature H_2S gas sensor based on ZnO nanoplatelets," Result in Physics, vol. 12, pp. 2189-2201, March 2019, doi:10.1016/j.rinp.2019.01.089.

[21] E. Espid, A. S. Noce, and F. Taghipour, "The effect of radiation parameters on the performance of photo-activated gas sensors," Journal of Photochemistry & Photobiology A, vol. 374, pp. 95–105, January 2019, doi:10.1016/j.jphotochem.2019.01.038.

[22] J. Cui, L. Shi, T. Xiea, D. Wang, and Y. Lin, "UV-light illumination room temperature HCHO gas-sensing mechanism of ZnO with different nanostructures," Sensors and Actuators B, vol. 227, pp. 220–226, 2016, doi:10.1016/j.snb.2015.12.010.

[23] E. Espid and F. Taghipour, "Development of highly sensitive ZnO/In_2O_3 composite gas sensor activated by UV-LED," Sensors and Actuators B, vol. 241, pp. 828–839, 2017, doi:10.1016/j.snb.2016.10.129.

[24] B. Gong et al., "UV irradiation-assisted ethanol detection operated by the gas sensorbased on ZnO nanowires/optical fiber hybrid structure," Sensors and Actuators B, vol. 245, pp. 821–827, 2017, doi:10.1016/j.snb.2017.01.187.

[25] A. Ilina et al., "UV effect on NO_2 sensing properties of nanocrystalline In_2O_3," Sensors and Actuators B, vol. 231, pp. 491–496, 2016, doi:10.1016/j.snb.2016.03.051.

[26] Z. Adamyan et al., "Nanocomposite sensors of propylene glycol, dimethylformamide and formaldehyde vapors," J. Sens. Sens. Syst., vol. 7, pp. 31-41, 2018, doi:10.5194/jsss-7-31-2018.

[27] C. Chen, G. Z. Li, J. H. Li, and Y. L. Liu, "One-step synthesis of 3D flower-like Zn_2SnO_4 hierarchical nanostructures and their gas sensing properties," Ceramics International, vol. 41, pp. 1857-1862, 2015, doi: 10.1016/j.ceramint.2014.09.136.

[28] C. H. Lin, S. J. Chang, W. S. Chen, and T. J. Hsueh, "Transparent ZnO-nanowire-based device for UV light detection and ethanol gas sensing on c-Si solar cell," RSC Advances, vol. 6, pp. 11146-11150, 2016.

[29] X. Xin et al., "UV-activated porous Zn_2SnO_4 nanofibers for selective ethanol sensing at low temperatures," Journal of Alloys and Compounds, vol. 780, pp. 228-236, 2019, doi: 10.1016/j.jallcom.2018.11.320.

[30] E. Wongrat, N. Chanlek, C. Chueaiarrom, B. Samransuksamer, N. Hongsith, and S. Choopun, "Low temperature ethanol response enhancement of ZnO nanostructures sensor decorated with gold nanoparticles exposed to UV illumination," Sensors and Actuators A, vol. 251, pp. 188–197, 2016, doi: 10.1016/j.sna.2016.10.022.

[31] S. H. Yan et al., "Synthesis of SnO_2-ZnO heterostructured nanofibers for enhanced ethanol gas-sensing performance," Sensors and Actuators B, vol. 221, 88–95, 2015, doi: 10.1016/j.snb.2015.06.104.

[32] R. Bajpai et al., "UV-assisted alcohol sensing using SnO_2 functionalized GaN nanowire devices," Sensors and Actuators B, vol. 171-172, pp. 499–507, 2012, doi: 10.1016/j.snb.2012.05.018.

[33] C. Zhao, J. Fu, Z. Zhang, and E. Xie, "Enhanced ethanol sensing performance of porous ultrathin NiO nanosheets with neck-connected networks," RSC Advances, vol. 3, pp. 4018–4023, 2013.

[34] L. Xu, W. Zeng, and Y. Li, "Synthesis of morphology and size-controllable SnO_2 hierarchical structures and their gas-sensing performance," Applied Surface Science, vol. 457, pp. 1064–1071, 2018, doi: 10.1016/j.apsusc.2018.07.018.

[35] C. Wang, L. Yin, L. Zhang, D. Xiang, and R. Gao, "Metal oxide gas sensors: sensitivity and influencing factors," Sensors, vol. 10, pp. 2088-2106, 2010, doi: 10.3390/s100302088.

[36] C. Zhang, G. Liu, X. Geng, K. Wu, and M. Debliquy, "Metal oxide semiconductors with highly concentrated oxygen vacancies for gas sensing materials: A review," Sensors and Actuators A, vol. 309, pp. 112026, July 2020, doi: 10.1016/j.sna.2020.112026.

[37] V. Brinzari and G. Korotcenkov, "Kinetic approach to receptor function in chemiresistive gas sensor modeling of tin dioxide. Steady state consideration," Sensors and Actuators B, vol. 259, pp. 443-454, April 2018, doi: 10.1016/j.snb.2017.12.023.

[38] R. Arora, U. Mandal, P. Sharma, and A. Srivastav, "Nano composite film Based on Conducting Polymer, SnO_2 and PVA," Materials Today: Proceedings, vol. 4, pp. 2733-2738, 2017, doi: 10.1016/j.matpr.2017.02.150.

[39] M. S. Aleksanyan, V. M. Arakelyan, V. M. Aroutiounian, and G. E. Shahnazaryan, "Investigation of Gas Sensor Made of In_2O_3:Ga_2O_3 Film," Journal of Contemporary Physics (Armenian Academy of Sciences), vol. 46, pp. 86-92, February 2011, doi: 10.3103/S1068337211020071.

[40] S. Zhanga, T. Lei, D. Li, G. Zhang, and C. Xie, "UV light activation of TiO_2 for sensing formaldehyde: How to be sensitive, recovering fast, and humidity less sensitive," Sensors and Actuators B, vol. 202, pp. 964–970, June 2014, doi: 10.1016/j.snb.2014.06.063.

[41] V. M. Arakelyan et al., "Gas sensors made of multiwall carbon nanotubes modified by tin dioxide," Journal of Contemporary Physics (Armenian Academy of Sciences), vol. 48, pp. 176-183, June 2013, doi: 10.3103/S1068337213040063.

[42] Z. Qina et al., "Highly sensitive alcohol sensor based on a single Er-doped In_2O_3 nanoribbon," Chemical Physics Letters, vol. 646, pp. 12-17, 2016, doi: 10.1007/s11664-008-0596.

[43] G. Feng, M. Zhang, S. Wang, C., Song, and J. Xiao, "Ultra-fast responding and recovering ethanol sensors based on CdS nanospheres doped with grapheme," Applied Surface Science, vol. 453, pp. 513-519, September 2018, doi: 10.1016/j.apsusc.2018.05.102.

[44] M. Aleksanyan, "Solid-State Sensors for Ethanol Detection," International Journal of Engineering and Artificial Intelligence (IJEAI), vol. 1, pp. 30-43, 2020.

Integrating Sensors and Virtual Reality for Volumetric CT Analyses of Agricultural Soil Samples

Leonardo C. Botega[1,2,3], Paulo E. Cruvinel[1,2]
[1]Embrapa Instrumentation, São Carlos, SP, Brazil
[2]Post-Graduation Programs in Computer Science - Federal University of São Carlos, SP, Brazil
[3]São Paulo State University, Marilia, SP, Brazil
Emails: leonardo.botega@unesp.br, paulo.cruvinel@embrapa.br

Abstract - **Multi-modal sensing techniques and data fusion from sensors can offer new possibilities for providing agricultural soil analysis in a robust manner. In this paper we report the results of integrating X-ray Tomography (CT), a non-invasive sensing technology, within a Virtual Reality (VR) environment for agricultural soils analyses. In such a context, through a user interface, sensors, and a volumetric visualization of tomographic images a set of agricultural soils samples has been submitted for porosity analyses. The use of graphic computational resources allowed the addition of functionalities, like volumetric visualization and immersion. For validation, it has been used a case study, involving analysis of porosity of agricultural soils samples. In fact, using energy of 59.6 keV and time window equal to 10 seconds for sampling of each tomographic projection it has been possible to reconstruct digital tomographic images from agricultural soils to be analyzed in such a system. Results indicate both the preferential paths for the water flow and a new way for evaluation of the physical properties of an agricultural soil.**

Keywords - X-ray sensors; virtual reality sensors; digital image processing; X-ray tomography; agricultural soil porosity; decision-making process.

I. INTRODUCTION

Direct and indirect measurements can be used to evaluate physical, chemical, and biological inputs availability in agricultural soils. In fact, both are based on the use of sensors. However, when there are needs for the spatial variability evaluation of those variables into agricultural soils, not only sensors but also methods should be taken into account. In fact, sensors and methods should be integrated to allow decision making related to the agricultural production processes [1].

Besides, evaluating the evolution that is happening in the soil science area, it is noticed an increasing interest of the scientific community in the development and application of non-invasive techniques for the study of physical characteristics of agricultural soils.

In such a context, since the 1980 decade, the application of image sensors based on Computed Tomography (CT) for agricultural soils imaging [2]–[9] has become one of the noninvasive methods for the evaluation of the water movement into soils due to morphology, and aggregates distribution. In fact, such kind of instrumental arrangement has provided improvements in relation to those techniques based on the use of gravimetric and neutron probe for water content measurements in agricultural soils [10][11].

Additionally, combined with the development of CT, new methods of three-dimensional (3-D) reconstruction were developed, mainly motivated by the lack of information from two-dimensional models for a precise diagnosis in studies that require volumetric information [12]. Another challenges regarding to such aspects were associated with the image reconstruction process, as well as those related to the reconstruction algorithms, the computational capacity, and the way to handle large amounts of data [13]. Therefore, since that time, it has been understood that working with tomographic reconstruction implied to take into account a large amounts of data and the need to have available a large processing capacity [14][15].

Moreover, due to the advent of precision agriculture, it had become imperative to have adequately models for management based on data analyses not only related to the spatial variability but also the temporal one in the areas used for agriculture.

In this sense, the standardization of data storage and the architecture of distributed information systems that allow integration of different types of data in a simple and transparent way had become to be quite important for the development of new methods for non-invasive analyses in agricultural industry [16]-[21].

Into such a subject, as an example, digital agricultural soil images are obtained by tomography taking into account several projections. Moreover, because one soil sample is scanned at different angles, a large amount of data should be computationally processed. Nowadays, the use of tomography not only allows us to obtain information about soil density and moisture at the pixel level but also allows quantification of the pore volume and its representation in three dimensions. The soil pores vary in size and shape and can be interconnected.

In 1982, Bouma has highlighted the importance to determine the continuity of the pore network for the flow of water in soil [22]. Therefore, not only pore diameter but also pore continuity interferes in the process of redistribution of soil water. In such a context, it is important to assess the porosity of the soil, because, depending on the soil management strategy adopted for planting, restriction of soil water flow may occur, thus compromising plant growth. To determine the soil porosity, volumetric measurements are conventionally used [23][24]. For this, it is necessary to collect undisturbed soil samples for quantitative evaluation of its porosity based on the use of tomographic scanners.

Besides, methods based on volumetric reconstruction have been developed for such a purpose, mainly due to the inadequacy of information of two-dimensional models for accurate diagnosis, in studies that need volumetric information. Thus, such methods suggest the composition of surfaces and volume of the samples under analyses, i.e., contributing with the increase of precision in process of information extraction. However, it is still a challenge gathering all the information from the agricultural soils, i.e., the continuity, size, and shapes of the pores in a soil sample, among others.

CT is one methodology that allows observing the structural components of the soil, allowing better visualization of the behavior of the structure and soil porous space. A bi-dimensional CT image indicates the amount of radiation absorbed by each portion of an analyzed sample. In fact, the amount of the radiation absorption can be associated to a calibrated scale.

Since the X-ray absorption capacity of a material is closely related to its density, different density areas can be represented by either pseudo-colors or by a gray tone values. Therefore, based on the intensity emitted by an x-ray source and the intensity captured by the detector at the other side of the propagation line, one can determine the attenuation weight due to the object that is located between the source and the detector. The data related with the attenuations and their weights are crucial for the reconstruction process from projections (Figure 1), which allows mapping all the linear attenuation coefficients into a slice of the sample.

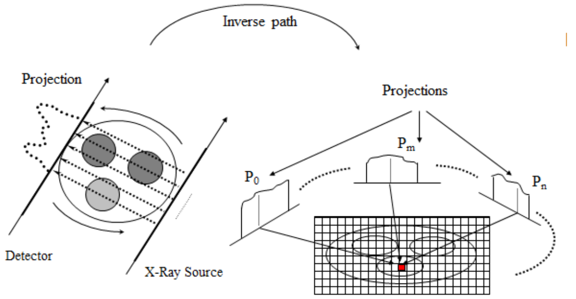


Figure 1. General view of the CT image reconstruction from projections
.

The calculation of the attenuated photons intensity in relation of the initial photons intensity can be obtained as follows:

$$N = N_0 e^{\oint_p (-\mu(\rho, Z_N, E)x)dp} \tag{1}$$

where (N) is the attenuated photons number, (N_o) is the initial photons intensity, (μ) is the linear attenuation coefficient (in cm^{-1}), (ρ) is the material density (g/cm^3), (Z_N) is the atomic number of the material, and (E) is the X-ray energy.

In addition, if the study sample is a chemical component or a mixture, like an agricultural soil, then its mass attenuation coefficient can be roughly evaluated based on the linear attenuation coefficients of each element. Furthermore, the final mass attenuation coefficients can be mapping, i.e., taking into account the spatial variability of the pixels, whose intensities can be given by:

$$\frac{\mu}{\rho} = \sum_i w_i \left(\frac{\mu_i}{\rho_i} \right) \tag{2}$$

where (w_i) is proportional to the weight of the i_{th} constituent of the sample`s material.

However, the interconnection for preferential flow requires additional methods, which can be beyond its use.

Besides, such an innovation can be faced by taking into account the composition of CT with sensors-based VR techniques, to assist noninvasive research through immersive and interactive processes.

The VR was born in the eighties under the need of differentiating traditional computational simulations of the synthetic worlds that began to stand out. This initiative gave credit to researchers like Bolt [25] and Lanier [26]. VR transports the individual into a fully immersive and interactive experience with a degree of realism. Likewise, academics, software developers and researchers have been still looking for defining a VR based on their own experiences. However, it is possible to observe in specialized literature that all of them technically considered the term related to the immersive and interactive experience, i.e., based on images generated by computers, rendering or not in real time [27]-[30]. Furthermore, the use of sensors in their external devices, i.e., digital gloves, video-helmets, digital caves, digital tables, among other, led to the concept related to sensors-based VR.

In 1994, Machover has stated that the quality of a VR system is essential, because it stimulates to the maximum the user, in a creative and productive way, providing feedbacks in a coherent way to the user’s movements [31].

Until the present moment, just some units of research have been developing projects using sensors-based VR applications in the area of scientific visualization, as the tomographic reconstruction, due to the high cost and to technical difficulties involved in such processes. However, some proposals have been appearing to minimize the difficulties of development and maintenance of the systems and necessary programs.

Additionally, today a better organization of human resources is being observed to integrate areas of the knowledge leading to the application of such advanced methods based on the connection and use of those technologies.

Thereby, the main objective of this work is to present the development of a VR system to support the analysis of 3D reconstructed soil samples using innovative immersive visualization and interaction techniques by integrating sophisticated external sensor-based devices.

Specifically, it is presented the organization and implementation of a synthetic environment, which makes possible the visualization, analysis and manipulation of soil samples produced by an algorithm of volumetric reconstruction of X-ray tomographic images, through graphic computational tools and non-conventional sensor based-VR devices, aiming immersion and user interaction to the scene entities, making possible the non-destructive analysis of agricultural soil samples, as shown in a case study in Soil Science.

The remainder of the paper is organized as follows: Section II presents the materials and methods; Section III presents the results, discussions, and performance evaluation; finally, conclusion and future work are presented in Section IV,

II. Materials and Methods

The conceptual and methodological structuring applied in the development of the sensors-based VR system dedicated to the inspection of digital tomographic images from agricultural soils, uses data obtained by means of a volumetric reconstruction algorithm. Figure 2 shows a general view of the sensors-based VR system dedicated to the tomographic inspection of agricultural soil samples, as well as the dataflow, where, from tomographic image data, such soil samples can be reconstructed, imported and treated by several VR processes, i.e., focusing analyses related to the soil science area.

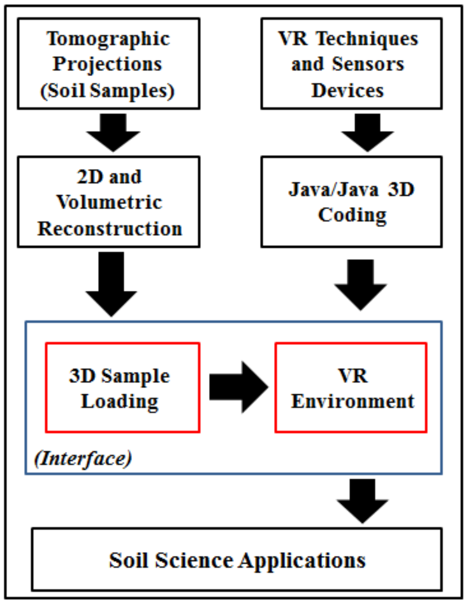

Tomographic Projections (Soil Samples)
VR Techniques and Sensors Devices
2D and Volumetric Reconstruction
Java/Java 3D Coding
3D Sample Loading
VR Environment
(Interface)
Soil Science Applications

Figure 2. General view of the sensors-based VR system customized to the inspection of tomographic samples of agricultural soils, as well as a view of the dataflow from acquisition to the visualization process.

The software system was organized taking into account the concept of classes. In object-oriented programming, a class is an extensible program-code-template for creating objects, providing initial values for state (member variables) and implementations of behavior (member functions or methods). In this work, the following classes have been considered: *Reconstruction*; *Loader*; *Transformations*; *Polygonal Attributes Extraction*; *Filter*; *Transparency*; *Illumination*; *Coloring*; *Conventional Collision*; *Non-conventional Collision*; *Conventional Model Manipulation*; *Non-conventional Model Manipulation*; *Conventional Scene Manipulation*; *Non-conventional Scene Manipulation*; *Quaternion*; *Visualization*; and *VR Environment*.

All devices were implemented using the Java programming language and the *Java3D* API [32].

For the obtaining of the tomographic image data, the CT scanner from Embrapa Instrumentation was used. All the tomographic projections allowed the images reconstruction, turning possible the generation of mass attenuation coefficient maps, i.e., given in [cm^2/g], with spatial resolution equals or larger than 1 mm. All the soil samples were submitted to the acquisition process under energy of 59.6 keV and time window equal to 10 seconds for sampling of the points for the tomographic projection.

For two-dimensional reconstruction, it was used an algorithm of *Filtered Back-Projection* (FBP), with a filtering based on the use of the *Hamming´s* window, implemented under 1-D *Fast Fourier Transform* (FFT), using the C++ language [33]. After that, with the 2-D reconstructed images, a suitable filtering technique was also used. Such a filtering technique was based on the use of *Wavelet Daubechies Transform* (WDT), which allowed filtering only certain image areas preserving borders and details, i.e., through using a window with 76 coefficients [34].

For the volumetric reconstruction it was adopted an interpolation based-overlapping algorithm of reconstructed two-dimensional slices. Such a technique consists of setting up the plans generated by the functions f(x, y, z_i) for i = 0... (n-1), where n is the number of reconstructed plans. Consequently, specific two-dimensional slices were interpolated to reconstitute the spaces left among these overlapped plans.

Figure 3 shows the original plans overlapping and the interpolated plans. Such a method was used to reduce the computational costs and the radiation time, based on the use of interpolation in between the spaces of the reconstructed slices based on the use of *B-splines* [35]. Thus, with only a few slices, the algorithm was prepared to estimate and complete the entire information.

Besides, the sensors-based VR system for the inspection of agricultural soils samples was organized taking into account the CT images, and a set of non-conventional sensors to support the *VR environment*.

In addition, for the evaluation of the preferential paths for the water movement in soil, sensors were used to detect the motion based on the use of gloves, the space based on 3-D visualization, i.e., using a CCD *head-mounted display*, as well as microelectromechanical actuators based on piezo-electrical devices [36][37].

Figure 4 shows photos of the used CCD glasses with sensors, model GSD 300 from Innovatek^{TM}, which has been used in the method for allowing the virtual reality including headsets sensors for properly align with the screen of the computational environment area, in order to reduce distortions.

Such sensors were necessary to translate the movement and to help the understanding of users in relation of the workspace into the agricultural soil samples. At the end of the process, the volumetric model is converted into *Wavefront File Format* (.obj), i.e., using the *vtkOBJExporter* class from *vtkOBJExporter.h* package of the visualization toolkit. This format has been chosen for its high performance and flexibility on import such models to virtual environment, where all their attributes can be customized for graphic API's.

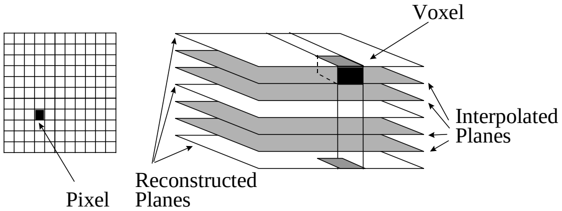
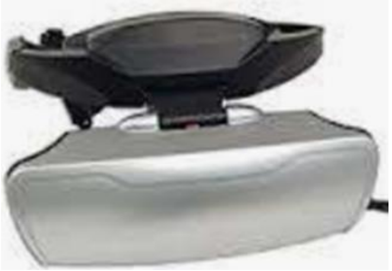
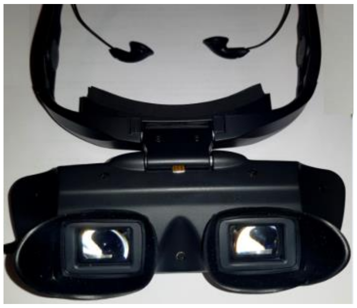


Figure 3. Volumetric reconstruction based on a set of reconstructed slices and the use of B-spline interpolator.



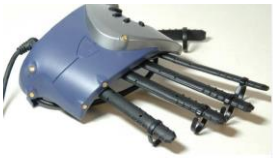
(a)



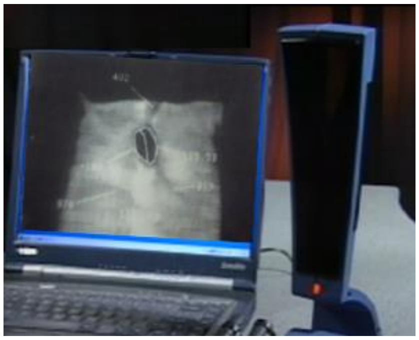
(b)

Figure 4. Details of the used virtual glasses based on sensors devices. In (a) the frontal mechanical view, and in (b) the CCD head-mounted display, microelectromechanical actuators, and headphones.

Figure 5 shows photos of the P5Glove with sensors obtained from the Mindflux^{TM}, which has been used for the 3-D virtual controller system. It is ergonomically adequate designed to allow for comfort during use. The glove features an infrared control receptor with an anti-reflective and scratch resistant lens. The main characteristics of the P5 glove includes a virtual 3D controller; Mouse-mode compatibility; 6 degrees of tracking (X, Y, Z, Yaw, Pitch and Roll) to ensure realistic movement; bend-sensor and optical-tracking technology to provide true-to-life mobility; as well as plug-and-play setup using an Universal Serial Bus (USB) port from a computational system.



(a)



(b)

Figure 5. (a) The P5Glove used for the 3-D virtual controller system; (b) the P5Glove`s control tower connected in a computer, i.e., based on infrared receptor with an anti-reflective lens.

The *Attributes Extraction* class treats of obtaining the voxels data from a volumetric image, using those above mentioned input non-conventional devices, i.e., supplying to the users' information on a specific point of the volumetric representation.

Initially, the objects of the classes *PickCanvas* and *PickResult* are instantiated, and these objects are responsible for activating the data extraction of a *Canvas3D* object and storing such data in vectors of events results.

Based on the user interest a region can be selected and attributes can be extracted using a coordinate z, since it can be stabilized on the selected region in the display, allowing selection through a two-dimensional viewport in an intuitive way.

Thus, the available data for picking operations under instances of Shape3D and their respective methods are: the borders, with *getBounds*; the scene graphs, with *getLocale* and *numBranchGraph*; the geometries, with *getGeometry*; *ColoringAttributes*, with *get.ColoringAttributes*; the material under the Hue, Saturation, Lightness (HSL) and Red, Green, Blue (RGB) formats, with *getMaterial*; the *transparency,getTransparency*; and the polygons, with the *getAppearance.getPolygonAttributes.getPolygonMod* class.

In addition, an object is instantiated, belonging to the *PickIntersection* class, also of the *com.sun.j3d.utils.picking* package, responsible for sheltering the collision point between an entity/node and the two-dimensional cursor. Thus, this instance stores in its content the intersection product among an entity of *PickResult* with the chosen Canvas3D point, which passed to the *getClosestIntersection* method as parameter. Like that, the *PickIntersection* class can offer through its events: the distance between the point and the observer, with the *getDistance* method; the coordinates of the point, with the *getCoordinates* method; the coordinates of the closest vertex, with the *getClosestVertexCoordinates* method; the normal straight line of the point, with the *getNormal* method; and the transformation head offices with the *getMatrix* method.

Besides, the classes *PickIntersection* and *PickResult*, as well as the *Attributes Extraction class* can allow the reading of each mass attenuation coefficient values, which are present in the tomographic volume. In this context, these values can be obtained through the gray level tones, which are represented by luminance, index "L" from the HSL pattern, obtained by using the class *getMaterial* method.

The *Non-conventional Scene Manipulation* class is one of the most important for the user interactivity and immersion in the *VR environment*, once it allows the user browsing in all directions through the synthetic scene, approximating and going into the reconstructed structures using the data gloves *P5Glove* class [38].

For the accomplishment of such events, the manipulation classes and the model of the scene are both based on another auxiliary class called *FPSGlove*, which is available in the *com.essentialreality* package offered by the devices manufacturer. The *FPSGlove* class is responsible for including all the parameters regarding the non-conventional devices, concerning to the positioning, orientation and fingers bending, making possible to detect the proximity and inclination on it, and thus launches a series of customized events.

On the other hand, in relation of the constructor method of classes, additional parameters of same importance can be activated, such as: (1) *P5_Init*; (2) *P5_setForward*; (3) *P5_setMouseState*; (4) *P5_setFilterAmount*; and (5) *P5_setRequiredAccuracy*. These classes are responsible for initializing, determining the positive direction, and turning off the mouse, filtering the sign and determining the precision movements, respectively. Soon afterwards, the methods responsible for detecting the position of the glove in the real environment are declared. The methods are the *getXPosition, getYPosition* and *getZPosition*, which map the triggers mentioned before to launch an event type, it means, they monitor the values received by the glove through instances of the class *P5State*, a class responsible for determining the current state of the glove. Thus, through the *filterPos* method of *P5State,* the exact position of the device is obtained and then assigned to the methods to check if the limits were or not outdated.

In a similar way to the positioning detection methods, still in the *FPSGlove* class, the methods *getYaw*, *getPitch* and *getRoll* are described, which answer for detecting the inclination of the device in the axis Y, X, and Z, determining if the established limits for the flags were reached.

After having implemented the monitors and triggers of events with the auxiliary class, the *Non-conventional Scene Manipulation* should now have their events described in its scope.

For such a scope, firstly the used class should be extended to the *ViewingPlatform* class, it means, to have their instances interpreted as events on the virtual scene. Besides, two other specific parameters are included, the translation step and the rotation one, which are responsible for defining when the virtual models will be moved or lean in each movement of the device in the real world, after being recognized by the *FPSGlove* class.

Once such a process is concluded, the instance of the *Non-conventional Scene Manipulation* class should be harnessed to the object of the ViewingPlatform class of the current *Canvas3D* object, so that all of the movements can be relationated on the scene and not the volumetric model, i.e., through the *setViewingPlatform* method.

The *Non-conventional Model Manipulation* is a class responsible for accomplishing the three-dimensional representation movements through real movements of the data glove *P5Glove*, where the user can change the positioning and orientation of models in real time in all directions and angles, contributing to the virtual reality environment interactivity in six degrees of freedom. To operate such a process, it is important to take into account the *Non-conventional Scene Manipulation*, in which the current implementation uses the support *FPSGlove* class.

Thus, the *Non-conventional Model Manipulation* class is extended of the Behavior class, a class that describes behaviors, customized for reactions to the movements of the device.

Besides, after assigning the procedures *getXPosition*, *getYPosition*, and *getZPosition* to obtain the positions, as well as the procedures *getYaw*, *getPitch*, and *getRoll* to obtain the orientations, under instances of the *FPSGlove* class, the procedure *rotateQuaternion* is called. Such an operation is based on the transformation of the Euler angles in Quaternion coordinates, i.e., useful for establishing the rotations using complex numbers and imaginary axis to improve the movements precision [39]. A Quaternion (q) is represented by:

$$q = \left(s, \vec{v}\right) \tag{3}$$

where the (s) and ($\vec{v}$) are representing the real and vectorial components respectively.

The relevance of using Quaternions is related with the opportunity for applying rotations in data collected from sensors, i.e., using three-dimensional models, based on the use of vectorial products. For the *Quaternium* class implementation the Java programming language has been used (Java3D API).

For the Quaternion`s model, we have divided the procedure in three main steps: (1) step for mapping the position and orientation of the glove, i.e., based on the data collected from positioning sensors, (2) step for coordinates conversion; and (3) step for the parameterization to allow rotation and visualization based on synthetic scenes entities.

The P5Glove used is an unconventional device having only 128 g. It is mouse compatible during operation. It also presents fold sensors, which are located on the fingers structure, i.e., being responsible for the identification of the movements, as well as the actions for holding a sample in the synthetic RV environment. Such sensors can have their parameters customized through the use of an appropriate API, i.e., called *Dualmode*. As observed previously in the technical features descriptions, its operation is based on an optical tracking system and two photosensitive receivers located in a mechanical tower. Also, to perform the positioning data mapping and orientation the glove has been used taking into account all the available six degrees of freedom. Likewise, an additional class has been developed to handle the signals from the glove`s hardware, interpreting the drivers provided by the manufacturer.

The coordinate's conversion takes place according to the Pseudocode 1 (Figure 6), which provides not only the scalar but also the vectorial part of the Quaternion. The Pseudocode 2 (Figure 7) presents the application of the Quaternion terms to the matrices transformation.

```
Begin
    $New rotation = \sin(angularStep/2)$
    $q_1 \cdot x = rotation * gloveAxis_1$
    $q_1 \cdot y = rotation * gloveAxis_2$
    $q_1 \cdot z = rotation * gloveAxis_3$
    $q_1 \cdot w = \cos(angularStep/2)$
End
```

Figure 6. Pseudocode 1, which is part of the code to convert from Euler to Quaternion.

```
Begin
    $transformationMatrix[0] = (1.0 - 2.0 * q_1 \cdot y * q_1 \cdot y - 2.0 * q_1 \cdot z * q_1 \cdot z) * scale[0]$
    $transformationMatrix[4] = (2.0 * (q_1 \cdot x * q_1 \cdot y + q_1 \cdot w * q_1 \cdot z)) * scale[0]$
    $transformationMatrix[8] = (2.0 * (q_1 \cdot x * q_1 \cdot z - q_1 \cdot w * q_1 \cdot y)) * scale[0]$
    $transformationMatrix[1] = (2.0 * (q_1 \cdot x * q_1 \cdot y + q_1 \cdot w * q_1 \cdot z)) * scale[1]$
    $transformationMatrix[5] = (1.0 - 2.0 * q_1 \cdot x * q_1 \cdot x - 2.0 * q_1 \cdot z * q_1 \cdot z) * scale[1]$
    $transformationMatrix[9] = (2.0 * (q_1 \cdot y * q_1 \cdot z - q_1 \cdot w * q_1 \cdot x)) * scale[1]$
    $transformationMatrix[2] = (2.0 * (q_1 \cdot x * q_1 \cdot z - q_1 \cdot w * q_1 \cdot y)) * scale[2]$
    $transformationMatrix[6] = (2.0 * (q_1 \cdot y * q_1 \cdot z + q_1 \cdot w * q_1 \cdot x)) * scale[2]$
    $transformationMatrix[10] = (1.0 - 2.0 * q_1 \cdot x * q_1 \cdot x - 2.0 * q_1 \cdot y * q_1 \cdot y) * scale[2]$
End
```

Figure 7. Pseudocode 2, which is part of the code dedicated to apply the Quaternion for the matrices transformation.

Hence, back to the main process, the *rotateQuaternion* procedure assigns to its class the axis and angles parameters, in Euler coordinates and returns a Quaternion description, a set used in the same *Quat4f* constructor, constructing a Quaternion of float, and after in setRotation executing a rotation with instances of *Transform3D*.

The *Quaternion* class implements a conversion algorithm so the system stops using just rotations on the axis x, y and z, and starts to accomplish orientations on some intermediate axis, defined by a vector that goes through the origin and reaches a point in the space. Such kind of an axis can be represented by a specific coordinate of the real device, e.g., the Cartesian coordinates (x, y, z) of one of the eight LED's present in the controller tower, which is used with the glove.

To accomplish this operation, imaginary bases and complex numbers are used, providing an alternative parameter for *setRotation*, method of the *Transform3D* class, which allows using a Quaternion as an argument. Thus, calling an instance of *Quaternion* to accomplish a rotation with the non-conventional device *P5Glove*, the orientation of the glove is interpreted by the *FPSGlove* class and translated by *Non-conventional Scene Manipulation* or *Model*, is converted from the Euler system to *Quaternion* base, returning the system new orientation coordinates, to be executed by the *Quat4f* method of *Transform3D*, which encapsulates the whole functioning of the Quaternion previously described.

At the end of the process, the product of *Non-conventional Model Manipulation* class is encapsulated in a *BranchGroup* object and assigned to the transformation group, *TransformGroup*, which conducts the three-dimensional representation movements, in a distinct way of the previous class. In such a way, not only all the movements' detection, but also the effective positioning change, and the entities orientation produce effects under the current models in the *Canvas3D* object.

The *Non-conventional Collision* class treats the implementation of a collision detection algorithm added to the *Non-conventional Scene Manipulation* class, restricted to events that use non-conventional input data devices, specifically the data glove *P5Glove*. In that way, through the algorithm, the users are also prevented to cross the faces of a three-dimensional representation during the browsing process in synthetic scenes, allowing only the cameras transpositions inside the empty spaces among such faces, simulating real physical processes.

Thus, each spatial position of the glove is tested as the current instance of itself, it means, to each direction of movements in some specific moment, where the possible alternatives are: left, right, up, down, forward, and back. After identified the positioning of the glove, in the moment of a supposed collision, the *Non-conventional Collision* class can blocks the device movements. Thus, the last movement of the glove when colliding is stopped, although the data glove can freely be moved in the real environment. This is caused by a new instantiation of the current positions of the glove, assigning to them, empty vectors, in other words, initialized in the origin, i.e., causing the immediate stop of the device movements.

Additionally, at the same time, when accomplishing any other move, which does not take them to a continuation of the blocking, the class interprets them and allows continuing the valid movements series, through a new instantiation of the mapped positions of the glove, using as parameter the position where the collision was begun and the linear step adopted by the class. At the end of this process, a *Shape3D* is added to *PhysicalBody* to detect in the browsing the scene being used by a user, allowing interaction and selection of each three-dimensional faces. Besides, the algorithm of such a class allows both preventing the browsing to continue or not in a scene, as well as an information of the direction of the glove movement, since active.

For the implementation of the *Visualization* class, the system interface prepares a volumetric tomographic image to be visualized. In such a way, the volumetric tomographic image is prepared to be adjusting to the 3-D model, i.e., occupying the whole extension of the *Canvas3D* object. Such activity contributes to improve the visualization quality, once the resolution of the HMD screens is inferior to the conventional LCD and CRT monitors.

III. RESULTS AND DISCUSSIONS

Based on the use of the tomographic projections and the two-dimensional reconstruction FPB algorithm it was possible to get volumetric images by means of the *B-spline* use. Figure 8 presents a set of examples related to the volumetric tomographic images obtained for stratified agricultural soil, degraded soil and a clay soil sample, respectively.

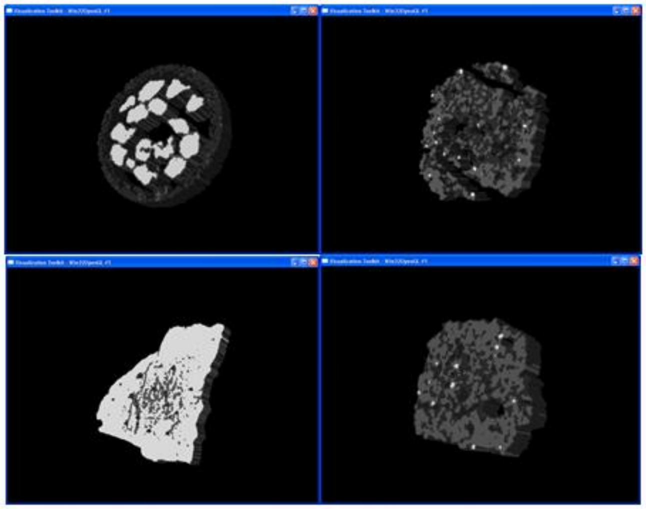


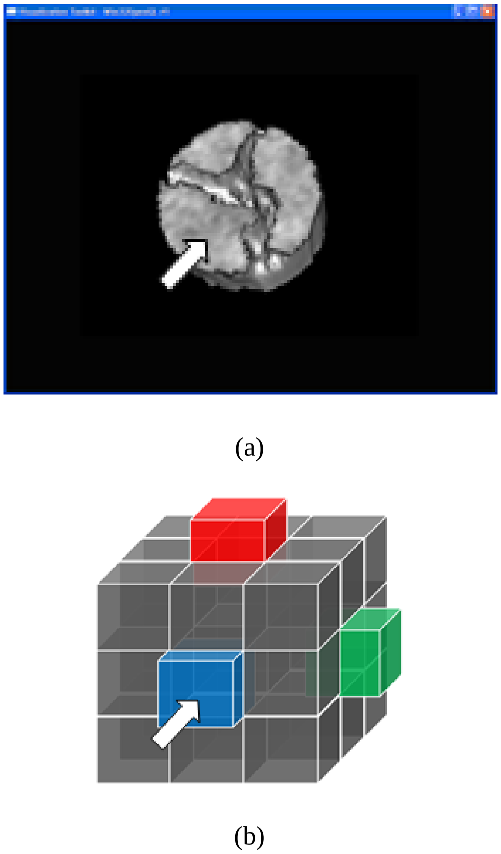
Figure 8. Volumetric images reconstructed by FBP, and the B-Spline algorithm

Based on the *Attributes Extraction* class, intrinsic characteristics of the scene and of agricultural samples could be obtained, through the use of either the mouse or the P5Glove, having as data origin the three-dimensional representations in the VR environment. Such a class allows the users to select any voxel from a set of voxels, which is useful for the 3D soil samples analysis during the information retrieval process. The set of data was divided into 2 categories: one of then concerning to the Scene, and other concerning to the CT measurements.

In relation to the first category, the synthetic scene data have been related to: borders, which have represented the geometry limits or the geometry limits that involved it; the scene graph, which has represented the node hierarchy in the tree; the current geometry in the model and its composition; the distance of a certain voxel in relation to the coordinates chosen in the scene; the closest vertex to the chosen point in the scene; the three-dimensional coordinates; and the normal straight line in the closest face, which had involved the chosen coordinates.

Secondly, concerning the tomographic data, the obtained data were: color attributes, which had represented the individual color of each voxel, independent of illumination intensity; the mass attenuation coefficients values of the agricultural soils, which are represented by the colors of each analyzed voxel, i.e., related to the light intensity in each position; transparency attributes; polygons attributes, and finally the saturation and matrix, of the HSL coefficients.

An experiment for validation of the result was prepared taking into account a digital and volumetric tomographic image obtained from a latosol soil. For such volumetric image, the value of an arbitrary voxel was taken as presented in Figure 9.



(a)

(b)

Figure 9. a) Volumetric image from a latosol sample, where an arbitrary point is chosen using a conventional device (mouse) or a non-conventional one (glove). The coordinates of the voxel are directly selected and respective information can be exhibited for users; b) Voxel selection from a 3D representation, i.e, illustrating the use of the *Attributes Extraction* class.

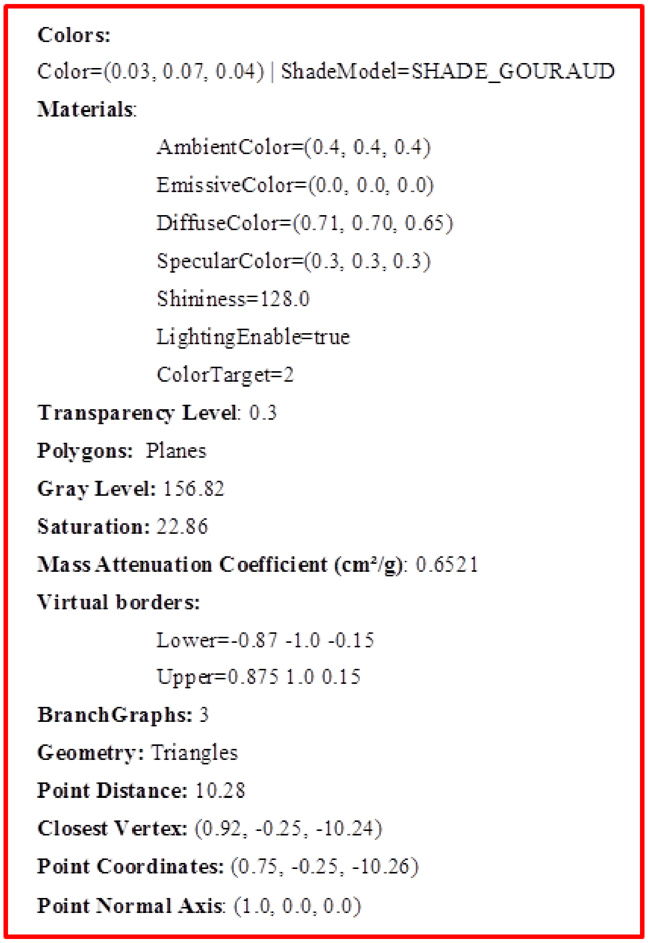
The developed method has allowed obtaining from a latosol samples sets of attributes, i.e., by assuring the reliable recovery of the samples agricultural data through the choice of voxels by selecting their coordinates (Figure 10).

In relation to the use of the *Non-conventional Scene Manipulation* Class, it has been possible to observe it usefulness to control the P5Glove, i.e., when one is browsing in a scene. In this context, according to this device position, users can be browsing through the scene, where the hand`s displacement can be faithfully translated to the scene`s movements in real time, even when the RV environment's cameras are moving. In an analogous way, such movements are also translated into the three-dimensional (3D) displacements. The *Non-conventional Model Manipulation* class simulates the manual support of 3-D samples, as well as its total movement inside the scene, with 6 degrees of freedom.

The *Quaternion* class has presented an adequate operation, making possible the correct conversion from the Euler`s coordinates to the Quaternion`s coordinates. Furthermore, the result obtained with the application of the Transform3D class has produced a smooth orientation changes for agricultural soil analysis.

Additional examples of results are presented below, i.e., considering an angular sector equal to 180º, divided into subsectors of 45º. For such examples, a set of voxels has been initially positioned at the coordinate (0.0, 0.0, 0.0).

Besides, the voxels have been rotated around an imaginal line, which has been represented by a dotted line. Moreover, it has been considered that such an imaginal line has both passed through the origin coordinate and reached the geographical position of three of the eight LEDs located on the control tower of the data relate to the glove.



Colors:

Color=(0.03, 0.07, 0.04) | ShadeModel=SHADE_GOURAUD

Materials:

AmbientColor=(0.4, 0.4, 0.4)

EmissiveColor=(0.0, 0.0, 0.0)

DiffuseColor=(0.71, 0.70, 0.65)

SpecularColor=(0.3, 0.3, 0.3)

Shininess=128.0

LightingEnable=true

ColorTarget=2

Transparency Level: 0.3

Polygons: Planes

Gray Level: 156.82

Saturation: 22.86

Mass Attenuation Coefficient (cm²/g): 0.6521

Virtual borders:

Lower=-0.87 -1.0 -0.15

Upper=0.875 1.0 0.15

BranchGraphs: 3

Geometry: Triangles

Point Distance: 10.28

Closest Vertex: (0.92, -0.25, -10.24)

Point Coordinates: (0.75, -0.25, -10.26)

Point Normal Axis: (1.0, 0.0, 0.0)

Figure 10. Sets of attributes and data from voxels by selecting their coordinates.

Table I, as well as Table II, and Table III, are showing results related to the intermediate rotations that are respectively associated with the following operations:

i. Dotted line segment, between the coordinates (-1.0, 1.0, 0.0) from the tower of LEDs and (0.0, 0.0, 0.0) from the origin (Figure 11);

ii. Dotted line segment, between the coordinates (1.0, 1.0, 0.0) from the tower of LEDs and (0.0,0.0,0.0) from the origin (Figure 12);

iii. Dotted line segment, between the coordinates (0.0, 1.0, 0.0) from the tower of LEDs and (0.0, 0.0, 0.0) from the origin (Figure 13).

TABLE I. RESULT OF THE ROTATION OF 180º, CONSIDERING THE GEOGRAPHICAL POSITION OF THE LEDS IN THE COORDINATE (-1.0, 1.0, 0.0).

Sample's Initial Position	Quaternions	Sample's Final Position
(0.0, 0.0, 0.0)	vector: (0.0, 0.0, 0.0), scalar: 0.92	(-0.70, 0.70, 0.0)
(-0.70, 0.70,0.0)	vector: (-0.27, 0.27, 0.0), scalar: 0.92	(-0.99, 0.99, 0.0)
(-0.99, 0.99, 0.0)	vector: (-0.38, 0.38, 0.0), scalar: 0.92	(-1.29, 1.29, 0.0)
(-1.29, 1.29, 0.0)	vector: (-0,49, 0,49, 0.0), scalar: 0.92	(-1.68, 1.68, 0.0)

TABLE II. RESULT OF THE ROTATION OF 180º, CONSIDERING THE GEOGRAPHICAL POSITION OF THE LEDS IN THE COORDINATE (1.0, 1.0, 0.0).

Sample's Initial Position	Quaternions	Sample's Final Position
(0.0, 0.0, 0.0)	vector: (0.0, 0.0, 0.0), scalar: 0.92	(0.70, 0.70, 0.0)
(0.70, 0.70, 0.0)	vector: (0.27, 0.27, 0.0), scalar: 0.92	(0.99, 0.99, 0.0)
(0.99, 0.99, 0.0)	vector: (0.38, 0.38, 0.0), scalar: 0.92	(1.29, 1.29, 0.0)
(1.29, 1.29, 0.0)	vector: (0.49, 0.49, 0.0), scalar: 0.92	(1.68, 1.68, 0.0)

TABLE III. RESULT OF THE ROTATION OF 180º, CONSIDERING THE GEOGRAPHICAL POSITION OF THE LEDS IN THE COORDINATE (0.0, 1.0, 0.0).

Sample's Initial Position	Quaternions	Sample's Final Position
(0.0, 0.0, 0.0)	vector: (0.0, 0.0, 0.0), scalar: 0.92	(0.0, 0.70, 0.0)
(0.0, 0.70, 0.0)	vector: (0.0, 0.27, 0.0), scalar: 0.92	(0.0, 0.85, 0.0)
(0.0, 0.85, 0.0)	vector: (0.0, 0.32, 0.0), scalar: 0.92	(0.0, 0.92, 0.0)
(0.0, 0.92, 0.0)	vector: (0.0, 0.35, 0.0), scalar: 0.92	(0.0, 0.95, 0.0)

Such examples of results are related to the operations that have allowed finding the new Quaternions. For them, the real part is resulting of the calculation from the cosine of the selected rotation angle. The imaginary part allows the evaluation of the directions of these new Quaternions, i.e., in relation to the reference axes. In fact, to use the approach presented in the examples above, the angular rotation step should be smaller than 180º. Actually, this represented constrains to avoid failures in finding the direction of the Quaternion. Nevertheless, a reliable operation has been obtained including calculation not only for the direction but also for the orientation of the Quaternion, as shown in Figure 14. Furthermore, by doing such operation no losses in relation to the degrees of freedom have occurred.

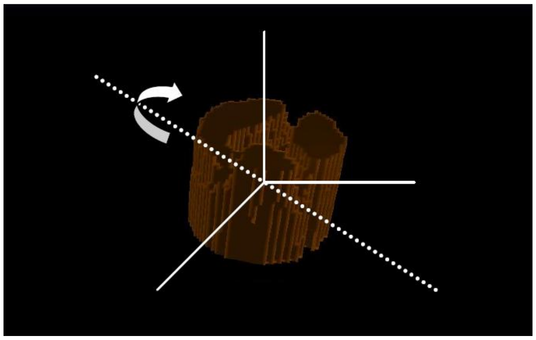


Figure 11. Result showing the rotation around the dotted line defined between the geographical position of the LEDs in the coordinate (-1.0, 1.0, 0.0) and the origin of the reference axes.

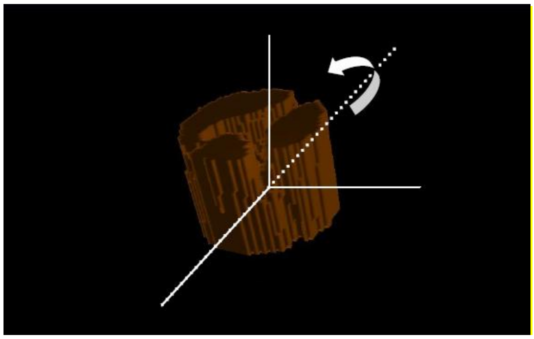


Figure 12. Result showing the rotation around the dotted line defined between the geographical position of the LEDs in the coordinate (1.0, 1.0, 0.0) and the origin of the reference axes.

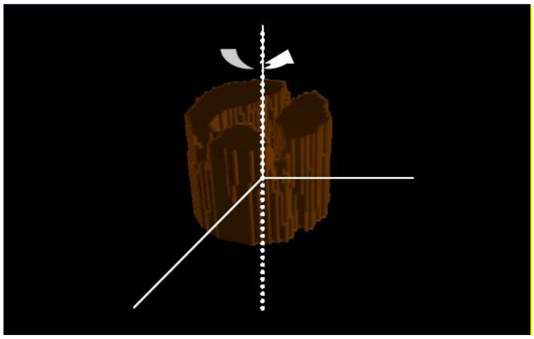


Figure 13. Result showing the rotation around the dotted line defined between the geographical position of the LEDs in the coordinate (0.0, 1.0, 0.0) and the origin of the reference axes.

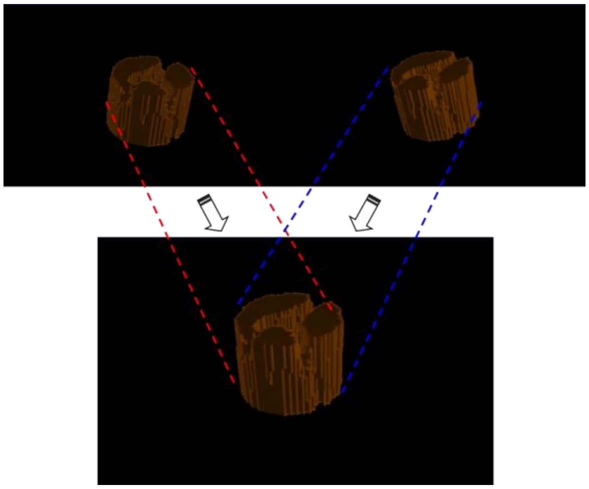


Figure 14. Representation of the rotation described around the stippled axis defined by the coordinate of LED = (1.0, 1.0, 0.0), and passing by the origin of the axes.

Through the *Visualization* class, the three-dimensional samples were examined by the Head Mounted Display in an immersive way. First, *Canvas3D*, responsible for the rendering of three-dimensional images, was maximized to omit the parts related to the main interface in the device, in order to focus only in the region where the sample was showed. Thus, each display of HMD forms an image, which is showing and interpreted by the user's brain with a larger depth effect.

Secondly, such an effect also has allowed performing the analyses of the preferential paths of the water flow into the agricultural soil samples, called as fingering effects, as well as the verification of the percentage of pores in the samples.

As described in the *Non-conventional Scene Manipulation* class, as the cameras are moved with the navigation processes, activated by keyboard interaction or data glove *P5Glove*, the traveled paths can be demarcated; leaving registered the itinerary, under visual and mathematical form.

Furthermore, for each device movement identified, it is established a new position for the camera, that means, given by new coordinates (x, y, z).

Such positions are unique and occupied only one per time. Thus, activated the demarcation process, from any point, the traveled path can be simulated for a certain water flow, i.e., when working with an agricultural sample. When accomplishing a certain movement, the current point occupied by the camera receives a *Shape3D* under the form of a blue sphere, which simulates the presence of a fluid drop occupying the previous position of the camera, leaving a bluish trace through where the camera passed.

In similar way to the simple scene manipulation, such demarcation obeys the laws imposed by the *Non-conventional Collision* class, it means, the traveled path is prevented of passing over the non-porous faces of the agricultural sample, leading the flow of fluids to pass among the related pores, the preferential paths.

The process can be repeated countless times, in way similar to the real situations.

Also, it is possible to make a borders calculation, where the limits of three-dimensional samples are identified in the space, as in the *Attributes Extraction* class, through the *getBounds* use on *Shape3D* instances, combined to a three-dimensional borders detection algorithm called *Polytope*, available in the Bounds package of *Java3D* API.

Such algorithm takes charge of drawing countless plans around of the surfaces of the sample, traveling all its extension in order to delimit exactly its borders. In such a way it allows that the nonporous parts of the samples, including the internal ones, can be identified, allowing the verification of its volume in cm³.

Figure 15 presents the result of the case study based on a tomographic image from a degraded agricultural soil, where the sample is in gray tones, and the water flow is represented with a blue color, demarcating the traveled paths.

In fact, once identified the non-porous part, the remaining portions were recognized based on the emptiness of the sample, which present the color that corresponds to those voxels in which occurred the absence of the photons attenuation. The porous voxel was filled out with a semi-transparent yellow color, seeking a larger prominence close to the sample. With such available data, it is possible to calculate the total volume of the sample (sum of the non-porous parts with its complement) in cm³.

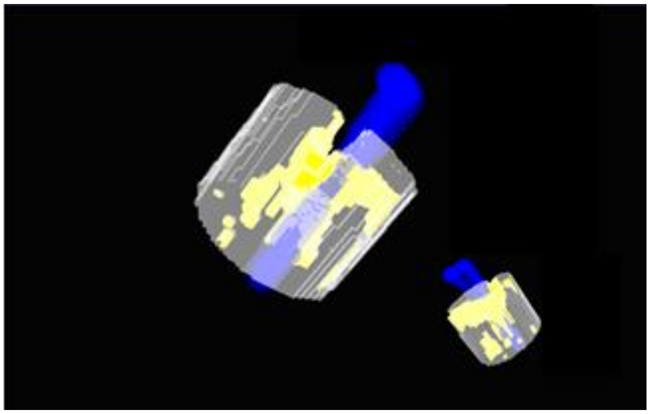


Figure 15. Result of the case study using a degraded soil sample, i.e., with representations of the non-porous soil portion (gray), the emptiness (yellow) and the water flow (blue) in between the soil pores.

Thus, starting from the total volume and the individual volume of the non-porous part, it is possible to calculate exactly the volume represented by the emptiness of the three-dimensional sample.

IV. CONCLUSIONS

This work presented the development of a new method, which took into account the integration of a sensors-based VR environment with a CT for dedicated inspection of agricultural soils. Results have shown both the possibility to access CT digital images from the agricultural soils, and the opportunity to handling three-dimensional manipulation and graphic visualization processes, through computational devices. Besides, such a developed method allowed the addition of the immersion and the user`s interaction with soil samples. Such resources had involved rendering control, illumination, coloring, attributes extraction and physical transformation, besides the integration of non-conventional data input and output devices, such as a *Head-Mounted Display* (video-helmet), and digital gloves.

In addition, it has been also observed that *Java3D* API provided in its group of classes, essential methods for HMD programming. Such development has encapsulated practically the entire stereoscopy programming.

Furthermore, the case study demonstrated the applicability of the method in visualization processes and agricultural soil sample analysis, considering the progresses and facilities when accomplishing non-invasive inspections.

Finally, the integration of CT and the sensors-based VR made possible to measure the volumes of emptiness of the samples, i.e., the pores, and simulate the water flow path for the formation of the preferential fingering. Future works will take into account embedded systems based on the use of *Field Programmable Gate Array* (FPGA), as well as use of the augmented reality concepts.

ACKNOWLEDGMENT

This research was partially supported by the São Paulo Research Foundation (FAPESP, Process No. 17/19350-2), and the Brazilian Corporation for Agricultural Research (Embrapa, Process No. 11.14.09.001.05.06). We thank the institutional support received from the Computer Science Department of the Federal University of São Carlos (UFSCar).

REFERENCES

[1] L. C. Botega and P. E. Cruvinel, “Sensors-Based Virtual Reality Environment for Volumetric CT Analyses of Agricultural Soils Samples”, *Proceedings of the IARIA, ALLSENSORS 2020: The Fifth International Conference on Advances in Sensors, Actuators, Metering and Sensing*, Valencia, Spain, 21-25 November 2020, pp. 27-34.

[2] A. Petrovic, J. Siebert, and P. Rieke, “Soil bulk density analysis in three dimensions by computed tomographic scanning”, Soil Science Society of America Journal, vol. 46, no. 3, pp. 445–450, 1982.

[3] J. M. Hainsworth and L. A. G. Aylmore, “The use of computer-assisted tomography to determine spatial distribution of soil water content”, Australian Journal of Soil Research, no. 21, pp. 435–440, 1983.

[4] S. Crestana, S. Mascarenhas, and R. Pozzi-Mucelli, “Static and dynamic threedimensional studies of water in soil using computed tomographic scanning”, Soil Science, vol. 140, no. 5, pp. 326–332, 1985.

[5] P. E. Cruvinel, R. Cesareo, S. Crestana, and S. Mascarenhas, “X-and gamma-rays computerized minitomograph scanner for soil science”, IEEE Transactions on Instrumentation and Measurement, vol. 39, no. 5, pp. 745–750, 1990.

[6] Á. Macedo et al., “Wood density determination by X and gamma ray tomography”, International Journal of the Biology, Chemistry, Physics and Technology of Wood, vol. 56, pp. 535–540, 2002.

[7] A. Pedrotti et al., “Computed tomography applied to studies of a planosoil” (Original in Portuguese: Tomografia compu-

tadorizada aplicada a estudos de um planossolo). Brazilian Agricultural Research Journal, vol. 38, no. 7, pp. 819–826, Brazil, 2003.

[8] P.E. Cruvinel, M. L. F. Pereira, J. H. Saito, and L.F. Costa, "Performance optimization of tomographic image reconstruction based on DSP processors", IEEE Transactions on Instrumentation and Measurement , vol. 58, pp. 3295-3304, 2009.

[9] J. M. Beraldo, F. A. Scanavinno Junior, and P. E. Cruvinel, "Application of X-ray computed tomography in the evaluation of soil porosity in soil management systems", Engenharia Agrícola, vol. 34, no. 6, pp. 1162–1174, 2014.

[10] E. S. B. Ferraz and R. S. Mansell, "Determining water content and bulk density of soil by gamma-ray attenuation methods", Technical Bulletin, no. 807, IFAS, Florida, pp. 1-51, 1979.

[11] C. F. A. Teixeira, S. O. Moraes, and M. A. Simonete, "Tensiometer, TDR and neutron probe performance in the determination of soil moisture and hydraulic conductivity", (Original in Portuguese: Desempenho do tensiômetro, TDR e sonda de nêutrons na determinação da umidade e condutividade hidráulica do solo), Brazilian Journal of Soil Science, vol. 29, pp. 161–168, 2005.

[12] M. F. L. Pereira and P. E. Cruvinel, "A model for soil computed tomography based on volumetric reconstruction, Wiener filtering and parallel processing", Computers and Electronics In Agriculture, vol. 111, pp. 151-163, 2015.

[13] K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and Optimization for Big Data Analytics", IEEE Signal Processing Magazine, pp. 18–31, 2014.

[14] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data", Knowledge-Based Systems, Elsevier, vol. 86, no.9, pp. 33–45, 2015.

[15] A. Ali, G. A. Shah, M. O. Farooq, and U. Ghani, "Technologies and challenges in developing machine-to-machine applications: A survey", Journal of Network and Computer Applications, vol. 83, pp. 124–139, 2017.

[16] S. S. Andrews, D. L. Karlen, and C. A. Cambardella, "The Soil Management Assessment Framework: A Quantitative Soil Quality Evaluation Method", Soil Science Society of America Journal, vol. 68, pp. 1945– 1962, 2004.

[17] A. Kaloxylos et al., "Farm management systems and the future internet era", Computers and Electronics in Agriculture, vol. 89, pp. 130– 144, 2012.

[18] U. Zimmermann et al., "A non-invasive plant-based probe for continuous monitoring of water stress in real time: a new tool for irrigation scheduling and deeper insight into drought and salinity stress physiology", Theoretical and Experimental Plant Physiology, vol. 25, no. 1, pp. 2-11, 2013.

[19] J. S. Selker, L. Graff, and T. Steenhuis, "Noninvasive time domain reflectometry moisture measurement probe", Soil Science Society of America Journal, vol. 57, no. 4, pp. 934-936, 1993.

[20] F. Palacios, M. P. Diago, and J. Tardaguila, "A non-invasive method based on computer vision for grapevine cluster compactness assessment using a mobile sensing platform under field conditions", Sensors, vol. 19, no. 17, pp. 3799-3818, 2019.

[21] H. Liu, R. Jia, X. Zhou, and L. Fu, "Virtual assembly of man-machine interactive mechanical seed-metering device based on matter-element identification", Transactions of the Chinese Society of Agricultural Engineering, vol. 32, no. 1, pp. 38-45, 2016.

[22] J. Bouma, "Measuring the conductivity of soil horizons with continuous macropores", Soil Science Society of America Journal, Madison, vol. 46, pp. 438-441, 1982.

[23] Y. Mualem, "A new model for predicting the hydraulic conductivity of unsaturated porous media", Water Resources Research, vol.12, pp. 2184-2193, 1976.

[24] M. Kutilek and D. R. Nielsen, Soil Hydrology, Cremlingen-Destedt: Catena Verlag, 1994.

[25] R. A. Bolt, "Put¬that¬there: Voice and gesture at the graphics interface", in 7th International Conference on Computer Graphics and Interactive Techniques, Washington, USA, pp. 262–270, 1980.

[26] J. Lanier, Visual programming languages, Scientific American, 1984.

[27] L. C. Botega and P. E. Cruvinel, "Development of a Virtual Reality Environment for Agricultural Soil Analysis" (Original in Portuguese: Desenvolvimento de Ambiente de Realidade Virtual para Análise de Solos Agrícolas), in Proceedings of the Workshop of Virtual and Augmented Reality, Itumbiara, Brazil, 2007.

[28] K. Pimentel and K. Teixeira, Virtual reality through the new looking glass, McGraw-Hill, New York, 2nd edition, 1995.

[29] O. Gonzalez et al., "Development and assessment of a tractor driving simulator with immersive virtual reality for training to avoid occupational hazards", Computers and Electronics in Agriculture, vol. 143, pp. 111-118, 2017.

[30] L. Jacobson, Garage Virtual Reality, SAMS Publication, Indianapolis, 1994.

[31] C. Machover and S. Tice, "Virtual Reality", IEEE Computer Graphics and Application, vol. 14, no.1, pp. 15-16, 1994.

[32] Sun Microsystems. Java3D Documentation. [Online]. Available from: http://java.sun.com/javase/technologies/desktop/java3d. Accessed December 2020.

[33] C. Kak and M Slaney, "Principles of computerized tomographic imaging," New york: The Institute of Electrical and Electronics Engineers, Inc., IEEE Press, 1988.

[34] I. Daubechies, "Ten lectures on wavelets", CBMS-NFS Regional Conference Series in Applied Mathematics, Philadelphia, PA: Society for Insdustrial and Apllied Mathematics (SIAM), vol. 61, 1992.

[35] T. N. E. Greville, "Spline functions, interpolation and numerical quadrature", Mathematical Methods for Digital Computers, Vol.2, A. Ralston and H.S. Wilf, eds., Wiley, New York, Ch. 8, pp. 156-168, 1967.

[36] S. Chen, L. Xu, and H. Li, "Research on 3D modeling in scene simulation based on Creator and 3dsmax," in IEEE International Conference, vol. 4, pp. 1736–1740, Mechatronics and Automation, 2005.

[37] E. F. S. Montero and D. J. Zanchet, "Virtual reality and medicine" (Original in Portuguese: Realidade virtual e a medicina), Brazilian Surgical Act, vol. 18, no. 8, pp. 489-490, 2003.

[38] C. Kenner. Essential reality P5glove sumary: Dual mode driver programming: [Online]. Available from: htpps//scratchpad.fandom.com/wiki/P5_Glove:Drivers_etc. Accessed December 2020.

[39] S. C. Biasi and M. Gattass. Use of Quaternions to represent 3-D rotations. (Original in Portuguese: Utilização de quatérnios para representação de rotações em 3-D), Catholic University of Rio de Janeiro, 2002. [Online]. Available from: http://www.tecgraf.puc-rio.br/~mgattass. Accessed December 2020.

CO_2 Detection by Barium Titanate Deposited by Drop Coating and Screen-Printing Methods

Fabien Le Pennec, Amine El Halabi, Sandrine Bernardini, Carine Perrin-Pellegrino, Khalifa Aguir,
and Marc Bendahan
Aix Marseille Univ, Univ Toulon, CNRS, IM2NP, Marseille, France
e-mail: fabien.lepennec@im2np.fr
e-mail: amine.elhalabi@im2np.fr
e-mail: sandrine.bernardini@im2np.fr
e-mail: carine.perrin-pellegrino@im2np.fr
e-mail: khalifa.aguir@im2np.fr
e-mail: marc.bendahan@im2np.fr

Abstract—**Metal Oxide Sensors are promising for gas detection but only a few studies about barium titanium deposition for carbon dioxide detection were reported. Its influence on detection has not been yet fully studied. Herein, we have realised barium titanium sensitive films by drop coating and screen-printing methods. A sensing material solution has been prepared by controlling the viscosity, and then the structural and morphological properties have been studied. The realised sensors were tested in the presence of CO_2 in dry and humid air (20%-50%-70%), in a concentration range from 100 ppm to 5000 ppm. Finally, a cross-interference study has been achieved with SO_2, NO_2 and CO interfering gases.**

Keywords-Gas Sensor; CO_2; $BaTiO_3$; Metal Oxide; Air Quality.

I. Introduction

Carbon dioxide (CO_2) is regularly studied as a target gas due to its wide involvement in many circumstances for security, health, or agricultural applications. Our previous work [1] has been focused on CO_2 sensing for the air quality control. CO_2 is present in the air we breathe. Its concentration in outdoor air is around 400 ppm [2]. It is an odorless, colorless, and non-flammable gas. Outdoor CO_2 emissions are mainly of natural origin such as volcanoes, and forest fires, or related to the breathing of animals and plants.

However, a small part of emissions (around a few %) comes from human activities, such as economic development [3], the energy sector (extraction of fossil fuels, electricity production, and heating provided by fossil fuel power plants) [4], agriculture (methane production) [5], industry [6], deforestation [6], transport, or buildings (construction, heating of residential and non-residential buildings) [7], [8]. CO_2 is a molecule also produced by the human body during respiration. Note that our respiratory and circulatory systems are sensitive to the CO_2 concentration. Indeed, an increase in the CO_2 concentration of the inspired air accelerates immediately our breathing rhythm. The CO_2 concentration inside buildings is usually between 350 and 2500 ppm and is related to human occupation and air renewal. Starting at 0.1%, CO_2 becomes a factor in asthma or building syndrome. At 4%, CO_2, the threshold for irreversible health effects is reached and a CO_2 level higher than 10%, can cause death. The CO_2 measurement can therefore be used as an indicator of air quality [9], [10].

Nowadays, the most commonly used CO_2 sensors are based on infrared phenomena, but this technology is expensive and miniaturization limited. Thus, the challenge of developing a CO_2 gas sensor with a good sensitivity, low-cost, which can provide reliable and reproducible detection results and a fast response to the target gas is increasingly claimed by different companies such as the environment, food industry, and medical. The electrochemical interaction of solid-state gas sensors meets these requirements. Indeed, many materials have been studied, in particular Metal Oxides (MOX) which have promising advantages as mentioned above [11], [12]. Iwata *et al.* [13] and Xiong *et al.* [14] worked on a CO_2 detector based on La_2O_3-SnO_2 and LaOCl-SnO_2, respectively. They obtained a high sensitivity of the sensor to a CO_2 exposure, besides Xiong *et al.* exhibit any saturation to a wide detection range (100 to 20 000 ppm). However, other materials have a high potential for CO_2 detection, such as the barium titanate (BaTiO3) presented in our previous work [1], whose semiconductor behavior is n-type. In 1991, Ishihara *et al.* [15] integrate $BaTiO_3$ in a mixed semiconducting oxide for CO_2 detection by a sensor based on a compressed disk. The combination of CuO-$BaTiO_3$, in equimolar proportion, bring a capacitive response equals to 2.98 for 2% of CO_2. A significant improvement in sensitivity has been achieved by adding silver to the composite. It has increased the sensor response up to 7.74 for 2% of CO_2 [16]. However, the operating temperature was still high (higher than 470°C) and a high concentration (20 000 ppm) was presented, which is not suitable for the air quality control application where the common concentration outdoor is 400 ppm and low energy consumption is required. In addition, M.-S. Lee *et al.* [17] and Keller *et al.* [18] have worked on another approach using a complex mixed semiconducting oxide, $BaTiO_3$-CuO-$LaCl_3$ and $BaTiO_3$-CuO-La_2O_3-$CaCO_3$, respectively. The latter was based on the deposition of a thick film, which was coated by a combination of laser ablation technique and screen printing. The study presents a sensitive layer with a response, R_{gas}/R_{air} = 2.8 for 5000 ppm of CO_2. Moreover, several publications tend to enhance the response to CO_2 through the development of thin films. For example,

TABLE I. SUMMARY OF CO_2 GAS SENSORS BASED ON A COMPOSITE OF $BaTiO_3$ GAS SENSOR

Sensing material	Depositing method	Response definition	Sensitivity	Temp. (°C) / R.H. (%)	Response / Recovery time	Refs.
$BaTiO_3$-CuO-$LaCl_3$	Screen printing	R_g/R_0	2.82 to 10000 ppm	550 / -	-	[17]
$BaTiO_3$-CuO-La_2O_3-$CaCO_3$	Screen printing	R_g/R_0	2.80 to 5000 ppm	600 / -	5 min / -	[18]
CuO-$BaTiO_3$-Ag	RF sputtering	R_g/R_0	1.22 to 5000 ppm	300 / 40	2 min / 3 min	[19]
CuO-$BaTiO_3$-Ag	RF sputtering	R_g/R_0	1.59 to 500 ppm	250 / -	1.5 min / 2 min	[22]
CuO-$BaTiO_3$-Ag	Brush coating	R_g/R_0	1.40 to 700 ppm	120	3 s / 5 s	[23]
$BaTiO_3$	Screen printing	R_g/R_0	1.71 to 400 ppm	280 / 50	2 min / 5 min	This work
$BaTiO_3$	Drop coating	R_g/R_0	1.73 to 400 ppm	280 / 50	2 min / 4 min	This work

Herrán *et al.* [19] carried out a study about $BaTiO_3$-CuO-Ag to improve CO_2 detection. Thus, compared to the use of a thick layer, the radio frequency (RF) sputtering method to obtain a thin metal oxide film brings many benefits, such as the sensitivity or the response/recovery time. However, the main advantage provided by the thin film is its influence on the sensitivity due to the contribution of the metal-semiconductor junction, which has led to a change in resistance [20]. Numerous studies [20, 21, 22] have also shown that it is possible to considerably improve the sensitivity of sensors based on $BaTiO_3$-CuO by the addition of metallic nanograins such as silver. Also, Joshi *et al.* [23] studied this composite and demonstrated a good sensitivity to CO_2 with long-term stability and excellent selectivity for low operating temperature (120°C). In the meantime, few authors report on a study on pure $BaTiO_3$. In Table I, we have presented a literature review of CO_2 sensors based on $BaTiO_3$. The methods of deposition of the sensitive films, the operating temperature, as well as the sensor performances, are summarized. S.B. Rudraswamy *et al.* [24] have shown that $BaTiO_3$ based on a thin film deposited by RF sputtering had a sensitivity to CO_2 equals to $R_{gas}/R_{air} = 1.1$ for 500 ppm. S. B. Rudraswamy *et al.* and B. Liao *et al.* have shown through their various studies [24], [25] that pure $BaTiO_3$ has no sensitivity to CO_2 in dry air. These observations can be explained by the need for the presence of moisture in the carrier gas mixed with CO_2 to obtain a change in the work function [26]. Therefore, as this material looks promising for CO_2 detection in wet conditions, we decided to manufacture a sensor using $BaTiO_3$ ink to develop sensors that are easy and inexpensive to manufacture.

In this paper, a comparison between the drop coating and the screen-printing methods are presented for the elaboration of $BaTiO_3$ low-cost thick film. The advantages of their use are the speed and the deposition simplicity. Thus, the electrical performances of $BaTiO_3$ during exposure to CO_2 are investigated. Both deposition methods are compared on the basis of several characteristics such as sensitivity, baseline stability, and response repeatability. The rest of the paper is structured as follows. In Section II, we describe our approach based on $BaTiO_3$ Nano-Powder (NP) deposition on platinum interdigitated electrodes by screen printing and drop coating, low cost, and easily used techniques. Then, in Section III, the detection results are discussed based on a change in the conductance of $BaTiO_3$ during the CO_2 introduction. The detection performances have been studied in a CO_2 concentration range between 100 and 5000 ppm, in the presence of humidity (R.H. 20% 50% and 70%). Finally, a conclusion is given in Section IV.

II. EXPERIMENTAL

This experimental section consists of two parts; in the first part, we have described the sensing film fabrication; in the second part, the measurement system set-up.

A. MOS gas sensors

To carry out our platform test, interdigitated Ti/Pt electrodes, 5 and 100 nm respectively, were deposited by Radio-Frequency (RF) magnetron sputtering on a Si/SiO_2 substrate (Fig. 1).

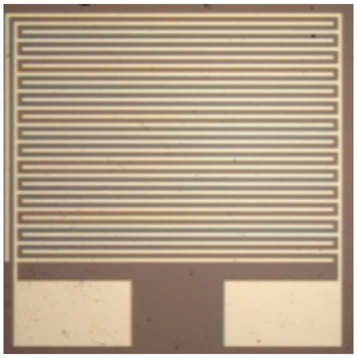


Figure 1. Transducer Ti/Pt interdigitated electrodes on a surface of 4x4 mm^2.

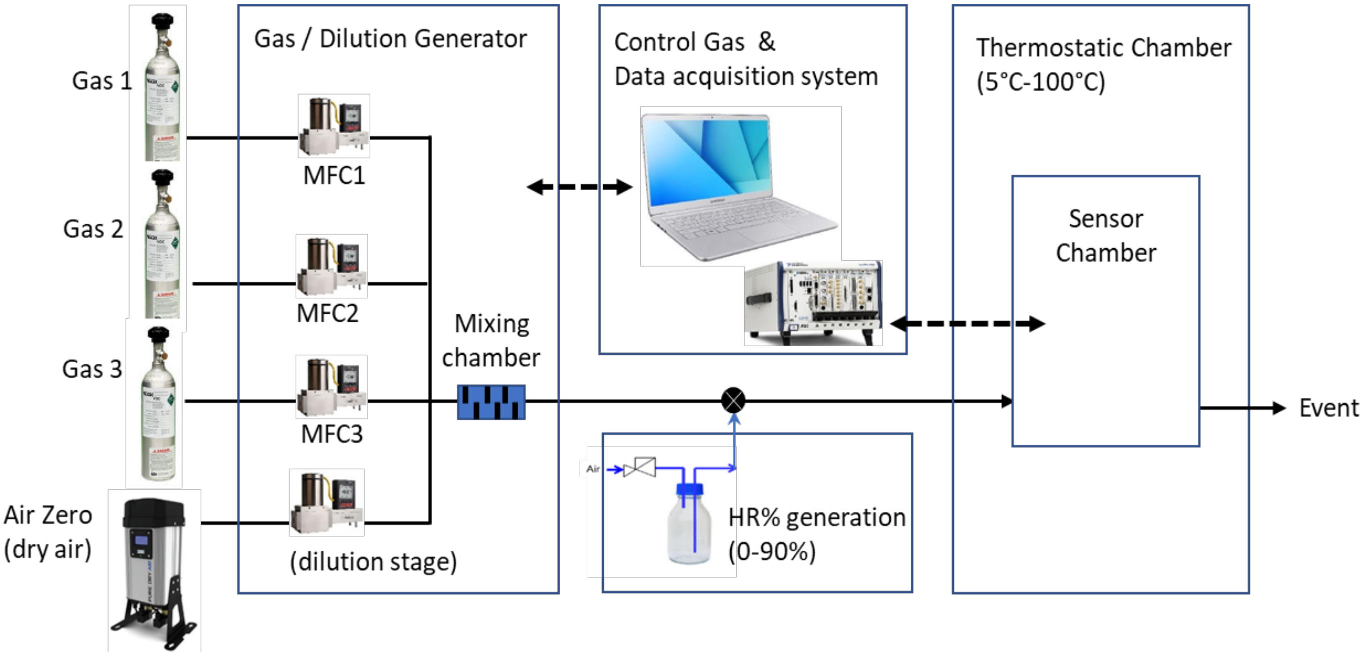


Figure 2. Description of the sensor test bench.

0.3g of $BaTiO_3$ Nano Particules (<100 nm, Sigma Aldrich®) was mixed with 0.3g of glycerol for the screen-printing. For the drop coating solution, 0.3g of $BaTiO_3$ NP was diluted in 5 mL of ethanol. Then, the solution was stirred by a magnetic agitator during 2h at room temperature. The solution viscosity has been adjusted to 2.7 mPa.s at 24°C with glycerol measured by the Sine-wave Vibro Viscometer SV-10 instrument. The solution was applied in drops on it using a glass Pasteur pipette. The gas sensors were annealed at 450 °C for 3 min in air ambient to evaporate the organic solvent and ensure the adhesion of the samples to the transducer. Then, the sensitive layer structure and the crystalline phase quality were checked by X-ray Diffraction (XRD) using an Empyrean Panalytical diffractometer equipped with a rapid detector with a theta-theta configuration and CuKα radiation (λ=0.154 nm). The surface investigation was performed by an SEM/EDS acquisition using a ZEISS GeminiSEM 500. Then, the thickness of the deposited $BaTiO_3$ films was measured with a surface profilometry mapping using a Bruker's DektakXT Stylus Profiler.

B. Electrical characterization

The test bench described in Fig. 2, consists of three parts, including a gas and a humidity generation system (0 to 90%), a thermostatically controlled chamber for regulating the temperature during the sensor characterization processes, and a data acquisition system. This equipment allows controlling the dilution of CO_2 in a carrier-neutral gas flow (air). Furthermore, the humidity is generated from the saturation flows of dry air by bubbling it in a container of deionized water, see Fig. 3. The humidity level is then regulated by measuring relative humidity with a capacitive probe and automatically controlling the mixing ratio using two mass flow controllers (MFC4 – MFC5) between the wet flow and the dry flow.

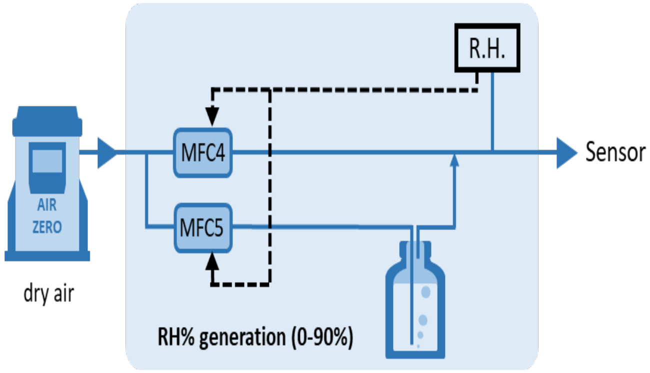


Figure 3. Description of the humidity generation in the air flow.

The gas dilution is precisely controlled by mass flow regulators. The thermostatically controlled chamber allows keeping the test chamber temperature constant during all the processes. The samples were located on a hotplate to control the operating temperature up to 300°C in a thermostatic chamber regulated at 30°C. The electrical measurements and the temperature were monitored by a homemade LabVIEW program to control the Source Measurement Unit (SMU) NI PXie-4141 and the programmable DC power supply NI-PXIe4113, respectively.

For the sensing property investigations, a 1V DC voltage was applied during the current measurements and a constant total flow was maintained by a MFC at 500 Standard Cubic Centimeters per Minute (SCCM). The CO_2 concentrations (from 100 to 5000 ppm) were generated by the mixture of synthetic air and the CO_2 diluted in dry air. Also, the CO_2 exposure was performed during 5 min with three Relative Humidity (RH equals to 20%, 50%, and 70%) to evaluate the sensor response. The sensors were operated at several temperatures from 200°C to 300°C. The best sensor performance compromise for this work was obtained at 280°C. The sensor response is defined in (1):

$$R = R_{gas} / R_{air} \qquad (1)$$

R_{gas} is the sensor resistance under CO_2 exposure and R_{air} is the sensor resistance in the air.

III. Results and Discussion

In this section, we will present the structural properties of the sensitive layer, and we will discuss our sensor performances.

A. *Structural characterization*

The XRD pattern presents in Fig. 4a the diffracted X-rays obtained with an Empyrean Panalytical diffractometer (λ=0.154 nm) at room temperature after deposited the screen-printing paste of $BaTiO_3$ on a Si/SiO_2 substrate and annealed it at 450°C during 3 min on a hotplate. Fig. 4b shows the diffracted X-rays obtained in the same conditions for the $BaTiO_3$ layer deposited by drop coating. The both diffractograms present a good agreement with the conventional tetragonal $BaTiO_3$ structure (PDF2 00-05-0626 (ICDD, 2002)) [27]. The $BaTiO_3$ layer deposited by screen printing presents some weak peaks visible in the diffractogram background that are not present in the $BaTiO_3$ layer deposited by drop coating indicating that this method leads to a layer with fewer impurities. For layers deposited by both techniques, the mean grain size was calculated to be 37 ± 2 nm using a single diffraction peak (111) and applying the Scherrer equation given by:

$$\tau = {}^{k * \lambda}/_{\beta \cos \Theta} \qquad (2)$$

where k = 0.9 and β is the peak FWHM (rad).

The (111) diffraction peak has been chosen as it is a single peak. Therefore, its width is supposed to depend only on grain size and instrumental width. However, the grain size calculation from one peak do not lead to an accurate estimation since it is representative to one preferential orientation.

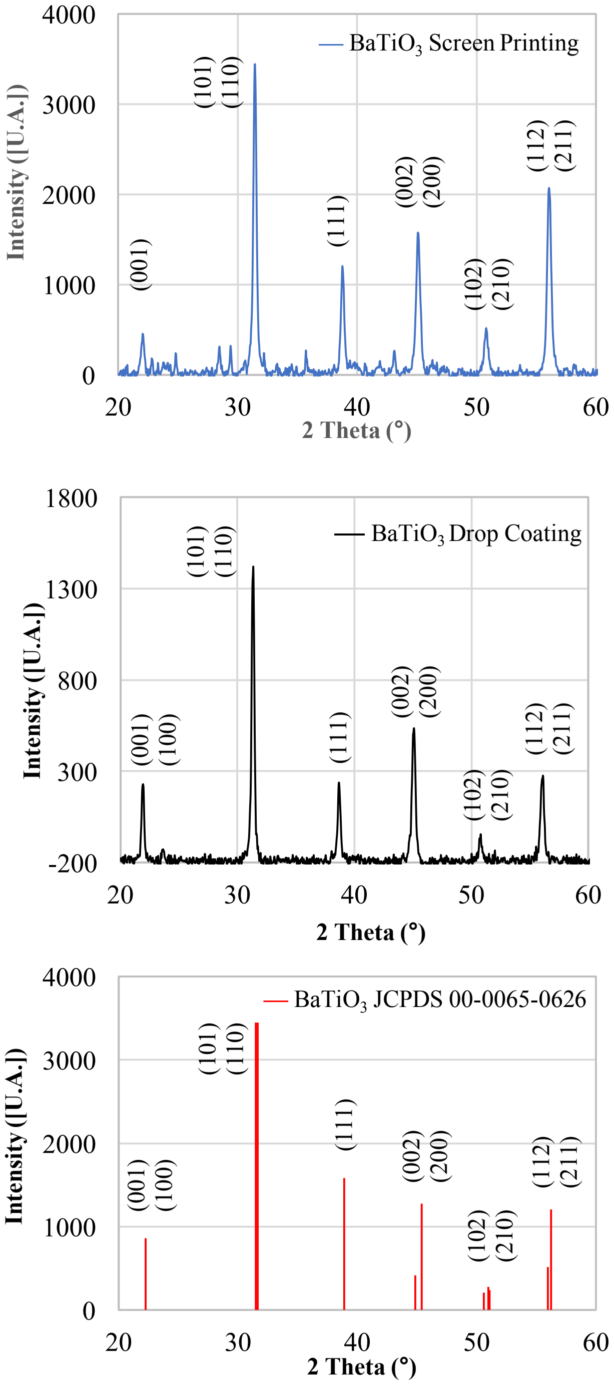


Figure 4. $BaTiO_3$ diffractograms for screen printing sensors (up) and drop coating ones (middle) using λ = 0.154 nm (Empyrean Panalytical equipment) compared with the reference pattern of the tetragonal structure (down) PDF2 00-05-0626 (ICDD, 2002) [27].

Thus, a scanning electron microscopy (SEM) image produced by a ZEISS GeminiSEM 500 (Fig. 5) enabled us to determine an average grain size which was estimated at 55 nm for both deposition methods.

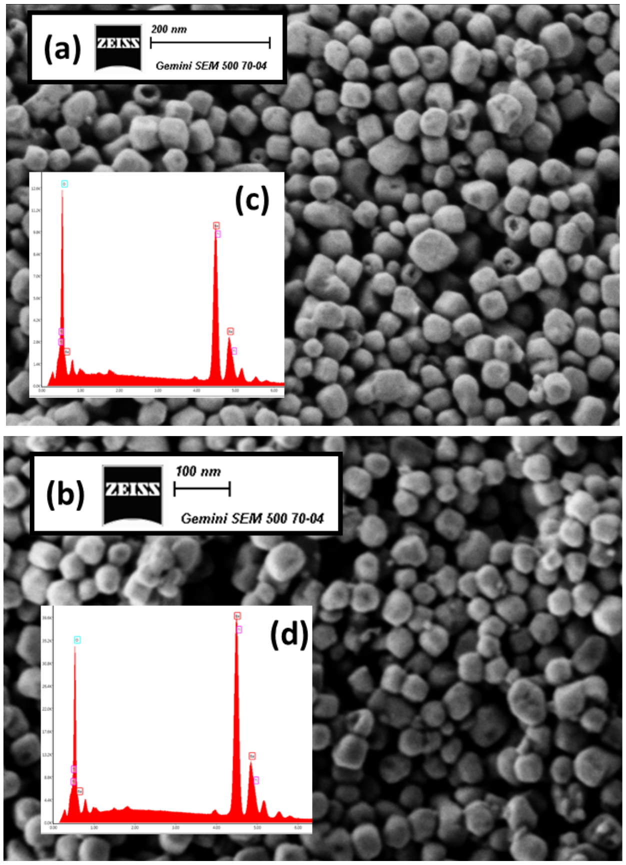


Figure 5. A SEM/EDS analysis have been performed for the both deposit method. SEM images of BaTiO3 (a) screen printing and (b) drop coating. EDS spectra of $BaTiO_3$ (inset) (c) screen printing and (d) drop coating.

The EDS spectra (inset (c) and (d) in Fig. 5) validate the stoichiometry of $BaTiO_3$ listed in Table II. The spectrums of the $BaTiO_3$ reveal the component of Barium (Ba), Titanium (Ti) and Oxygen (O). The EDS analysis is in agreement with the XRD analyses.

TABLE II. Comparison of Elemental Composition For Screen Printing and Drop Coating Obtained by EDS.

Deposit method	**Screen Printing**		**Drop coating**	
Element	**Atomic %**	**Weight %**	**Atomic %**	**Weight %**
Ba L	24	64.1	21.4	61.3
Ti K	19.7	18.4	18.7	18.7
O K	56.3	17.5	59.9	20.0

The surface morphological images (Fig. 6) were performed by a Bruker's DektakXT Stylus Profiler. The mean thicknesses, are estimated to be 30 µm and 15 µm for the screen printing and drop coating, respectively.

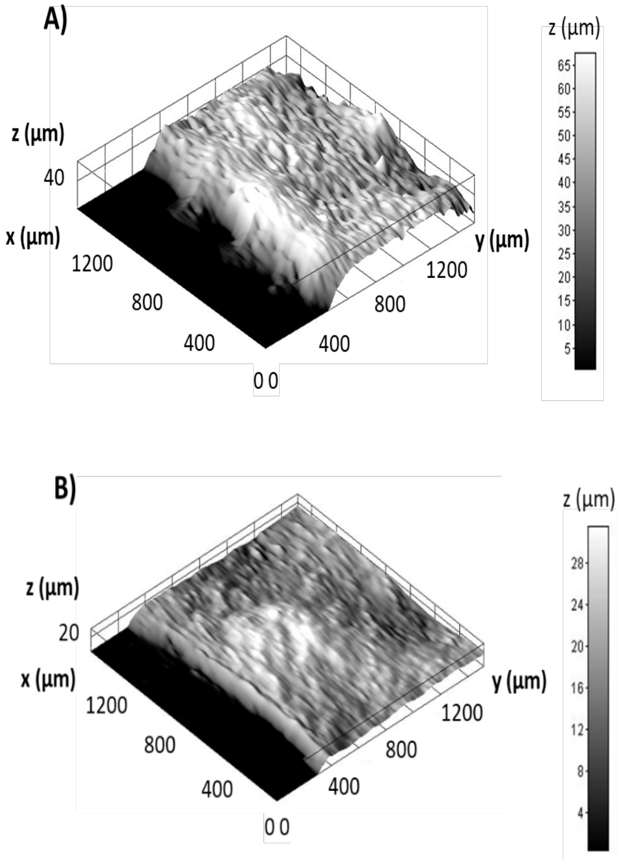


Figure 6. The surface morphological images: A) film surface deposit by screen printing and B) film surface deposit by drop coating.

The surface topography shows a high roughness of our $BaTiO_3$ layer for both deposition methods.

B. *Electrical sensor study for screen printing deposition method*

Fig. 7 shows a reversible response of the $BaTiO_3$ sensor to 400 ppm of CO_2 gas in 50% RH at 280°C. We observed the sensor resistance increase in the presence of CO_2. Since CO_2 is an oxidant gas, the sensor resistance increase confirms the n-type behaviour of $BaTiO_3$, according to [20]. The response and the recovery times are 2 minutes and 4 minutes, respectively. Where the response time τ_{res} is defined as the time required for the sensor to reach 90% of the sensor response, and the recovery time τ_{rec} as the time needed to reach 10% of the initial resistance baseline after the analyst gas has been purged.

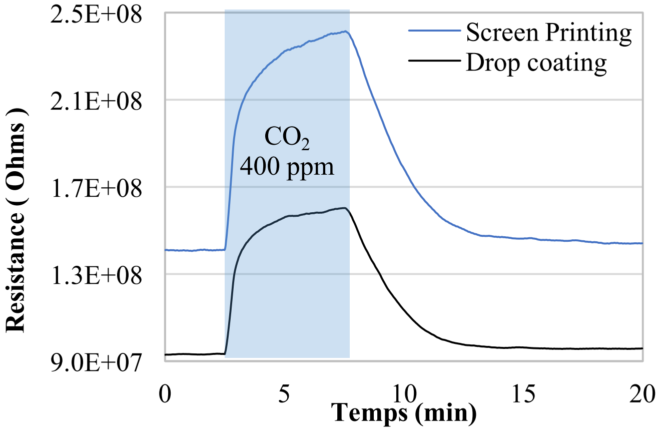


Figure 7. Resistance variation for 400 ppm of CO_2 at 280°C and 50% RH for sensors fabricated by screen printing (up) and drop coating (down).

By maintaining the same operating temperature of 280°C and 50% RH, the CO_2 sensor responses were measured from 100 ppm to 5000 ppm for both sensors and presented in Fig. 8.

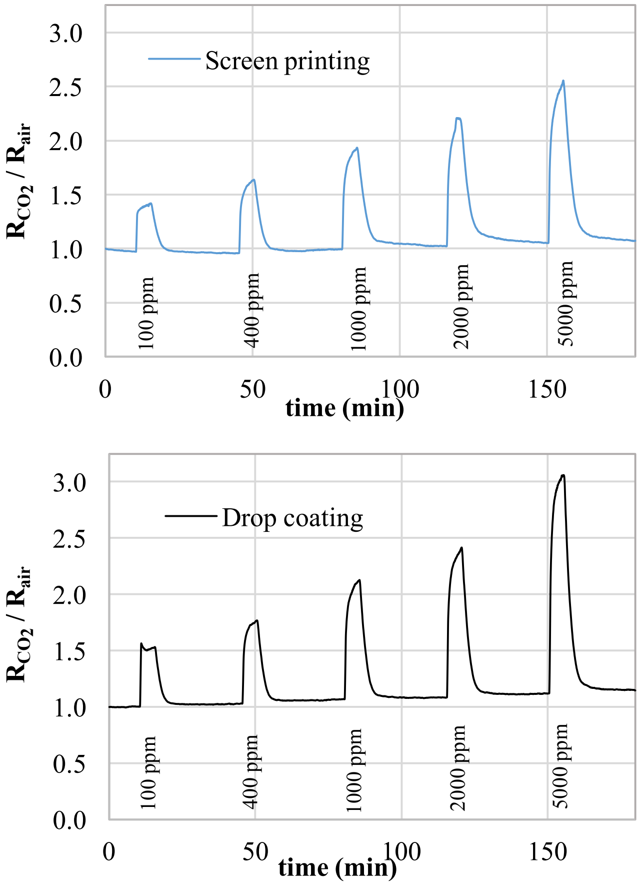


Figure 8. Sensor responses versus CO_2 concentrations (100-5000 ppm, 50% RH at T = 280°C) for screen printing and drop coating $BaTiO_3$ deposition.

In humid conditions and high temperatures, it is assumed that the CO_2 detection phenomenon follows the pathways indicated below [20], [26]:

$$BaTiO_3 + CO_2 \leftrightarrow BaCO_3 + TiO_2 \tag{3}$$

$$CO_3^{2-}|_{bulk} + CO_2|_{gas} + H_2O|_{gas} \leftrightarrow 2HCO_3|_{surface} \tag{4}$$

$$BaCO_3 + H_2O + CO_2 \leftrightarrow Ba(HCO_3)_2 \tag{5}$$

The gas sensors provide a measurable response to CO_2 as well as a stable baseline during the experiment. These results showed that our sensors have a wide detection range. It is possible to measure low concentrations with a low signal-to-noise ratio. Table III shows the comparison of the samples regarding their response and recovery times, respectively, tested from 100 to 5000 ppm at 280°C as operating temperature and 50% RH.

TABLE III. COMPARISON OF RESPONSE AND RECOVERY TIMES FOR SCREEN PRINTING AND DROP COATING

Screen printing		
CO_2 (ppm)	**response time (min)**	**recovery time (min)**
100	1.7	4.0
400	2.3	5.3
1000	2.8	6.3
2000	2.8	7.4
5000	3.0	6.0

Drop coating		
CO_2 (ppm)	**response time (min)**	**recovery time (min)**
100	1.5	4.0
400	2.3	4.2
1000	2.9	4.1
2000	2.4	4.6
5000	2.0	4.5

To study the humidity impact on the sensor responses, three levels of relative humidity (20%, 50 %, and 70%) were introduced into the test chamber and the recorded normalized responses, defined in (1), were evaluated (Fig. 9).

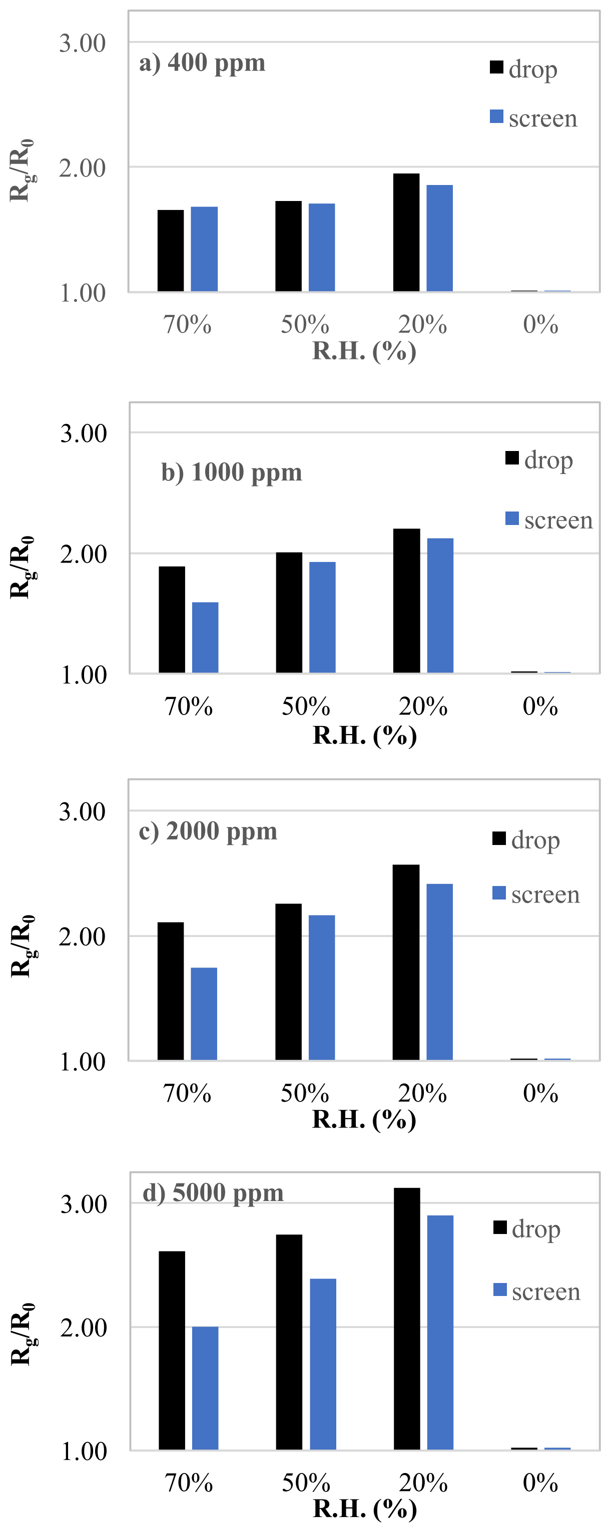


Figure 9. Normalized resistance of $BaTiO_3$ sensor upon CO_2 exposure at 280°C and with four relative humidity:
a) 400 ppm, b) 1000 ppm, c) 2000 ppm, d) 5000 ppm.

We have noticed that the humidity has an impact on the CO_2 response. For the different concentration levels, the sensor responses increase as humidity decrease. Therefore, optimal sensor responses were determined for 20% humidity. Table IV lists the CO_2 response values as a function of humidity and deposition method.

TABLE IV. A SUMMARY OF CO_2 RESPONSES BASED ON R_G/R_0 FROM FIG. 9

Screening printing			
CO_2 (ppm)	**70%**	**50%**	**20%**
100	1.50	1.46	1.62
400	1.68	1.71	1.86
1000	-	1.93	2.12
2000	1.74	2.16	2.41
5000	2.00	2.39	2.90

Drop coating			
CO_2 (ppm)	**70%**	**50%**	**20%**
100	1.47	1.53	1.64
400	1.66	1.73	1.95
1000	1.89	2.01	2.20
2000	2.11	2.26	2.57
5000	2.61	2.75	3.13

In addition to sensitivity, reproducibility was examined in another set of experiments. However, the repeatability characteristics of the sensors were obtained at 50% RH, which is the value commonly used in the industrial sector. These results are presented in Fig. 10 and show good reproducibility of conventionally prepared sensors. Furthermore, we calculated the coefficient of variation (Table V) to evaluate the repeatability features, defined in (3):

$$C_v = \mathrm{SD} \,/\, x_{moy} \qquad (3)$$

where SD is the standard deviation and x_{moy} the average of the normalized response for CO_2 exposure.

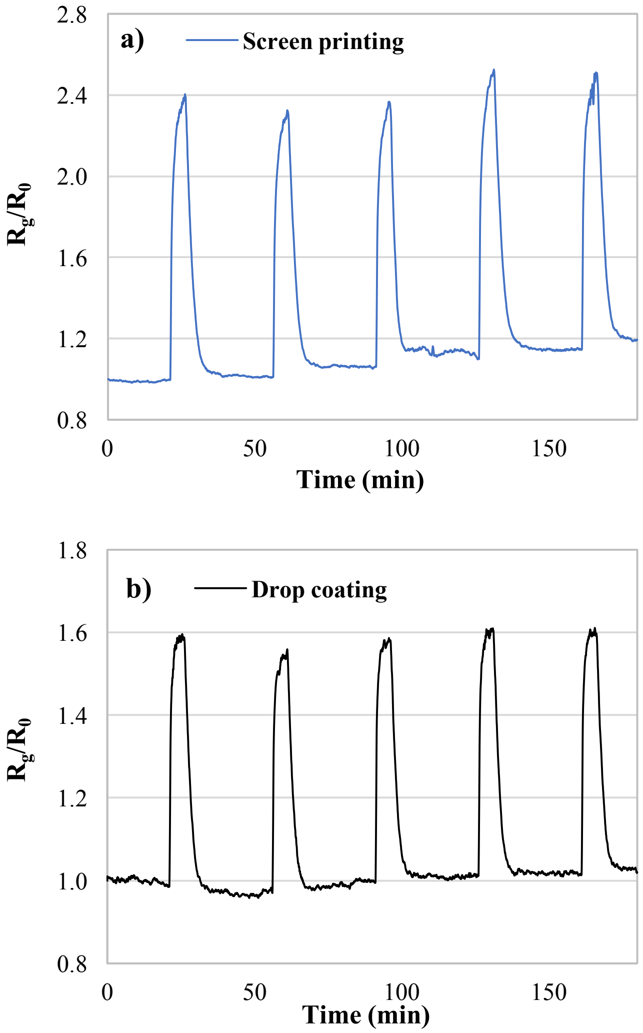


Figure 10. Normalized resistance of $BaTiO_3$ sensors for five exposures of 1500 ppm CO_2 at 280°C and 50% RH, a) screen printing and b) drop coating.

We determined a C_v equals to 2.84 % and 1.24 % for the sensors prepared by screen printing and drop coating, respectively. It indicates good repetition behaviour during each CO_2 exposure.

TABLE V. A SUMMARY OF CO_2 RESPONSES BASED ON R_G/R_0 FROM FIG.10

	R_g/R_0					C_v (%)
Screen printing	2.27	2.20	2.26	2.35	2.42	2.84
Drop coating	1.60	1.62	1.59	1.62	1.64	1.24

As MOX sensors are known for their poor selectivity, a cross sensitivity study of our $BaTiO_3$ sensors to three other greenhouse gases was carried out and is presented in Fig. 11.

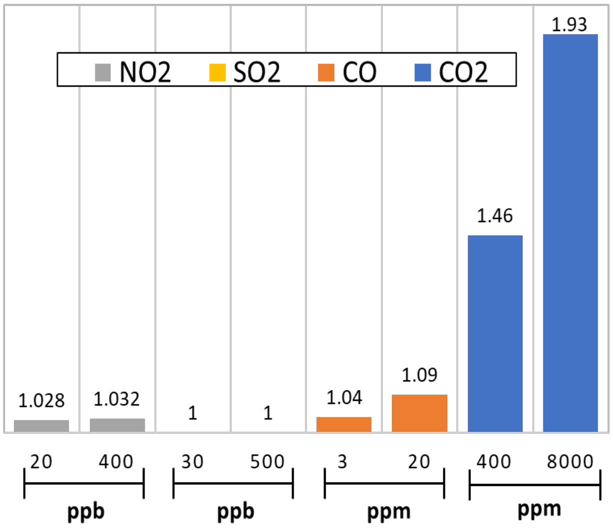


Figure 11. Selectivity study for four gases: NO_2, SO_2, CO, and CO2. Variation of the normalized resistance of $BaTiO_3$ sensors depending on the gas concentrations.

The gas concentrations chosen are based on the exposure limit value recommended by health agencies. This figure highlighted that our $BaTiO_3$ NP-based sensors have a high sensitivity to CO_2 compared to other gases.

IV. CONCLUSION

Metal Oxide are often studied to find the best materials to fabricate miniaturized and inexpensive sensors. However, the deposition method also influences the properties of the sensors. In this work, two methods of $BaTiO_3$ NP deposition were compared: screen printing and drop coating. The crystalline quality of the deposit was then checked for both sensor series. The sensitive layers formed by the $BaTiO_3$ material were tested as CO_2 sensors at an optimized temperature of 280°C and three relative humidity values. The CO_2 concentration is proportional to the increasing resistance of the sensitive layer and the sensor baselines are relatively stable during the experiment. Moreover, the sensor response increases with a lower level of humidity in the carried gases. The $BaTiO_3$ sensors have good repeatability feature to CO_2 exposure. For the sensors fabricated by screen printing, the response and the recovery times were determined to be 2 min 30 s and 6 min, respectively, and 2 min and 4 min for the sensors with droplet coating layers. This work demonstrates a slight improvement in the performances of CO_2 sensor with the drop coating method. This observation would be due to a better control of the homogeneity thickness of the sensitive layer.

ACKNOWLEDGMENT

The authors gratefully acknowledge Mr. S. MOINDJIE for his electronic support, Mr. M. BERTOGLIO for his technical support and Mrs A. CAMPOS for her microscopy support. This research was supported by the French government through Ph. D grant.

REFERENCES

[1] F. Le Pennec, S. Bernardini, M. Hijazi, C. Perrin-Pellegrino,K. Aguir, and M. Bendahan, "Screen Printed $BaTiO_3$ for CO_2 Gas Sensor", ALLSENSORS 2020: The Fifth International Conference on Advances in Sensors, Actuators, Metering and Sensing, Copyright (c) IARIA, pp. 24-25, November 2020. ISBN: 978-1-61208-766-5 7

[2] J. Hansen, M. Sato, P. Kharecha, D. Beerling, V. Masson-Delmotte, M. Pagani, M. Raymo, D. L. Royer, and J. C. Zachos, "Target Atmospheric CO_2: Where Should Humanity Aim? ", The Open Atmospheric Science Journal 2(1), pp. 20, May 2008, doi: 10.2174/1874282300802010217

[3] X.-P. Zhang and X.-M. Cheng, "Energy consumption, carbon emissions, and economic growth in China", Ecol. Econ., vol. 68, n^o 10, pp. 2706‑2712, August 2009, doi: 10.1016/j.ecolecon.2009.05.011.

[4] Danish, M. A. Baloch, N. Mahmood, and J. W. Zhang, "Effect of natural resources, renewable energy and economic development on CO_2 emissions in BRICS countries", Sci. Total Environ., vol. 678, pp. 632-638, August 2019, doi: 10.1016/j.scitotenv.2019.05.028.

[5] N. H. A. M. Ridzuan, N. F. Marwan, N. Khalid, M. H. Ali, and M.-L. Tseng, "Effects of agriculture, renewable energy, and economic growth on carbon dioxide emissions: Evidence of the environmental Kuznets curve", Resour. Conserv. Recycl., vol. 160, pp. 104879, September 2020, doi: 10.1016/j.resconrec.2020.104879.

[6] A. Shakoor, F. Ashraf, S. Shakoor, A. Mustafa, A. Rehman, and M. M. Altaf, "Biogeochemical transformation of greenhouse gas emissions from terrestrial to atmospheric environment and potential feedback to climate forcing", Environ. Sci. Pollut. Res., August 2020, doi: 10.1007/s11356-020-10151-1.

[7] D. J. Wales, Julien Grand, V. P. Ting, R. D. Burke, K. J. Edler, C. R. Bowen, S. Mintova, and A. D. Burrows, " Gas sensing using porous materials for automotive applications ", Chem. Soc. Rev., vol. 44, n^o 13, p. 4290‑4321, July 2015, doi: 10.1039/C5CS00040H.

[8] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information", Energy Build., vol. 40, n^o 3, p. 394‑398, January 2008, doi: 10.1016/j.enbuild.2007.03.007.

[9] A. P. Jones, "Indoor air quality and health", Atmos. Environ., vol. 33, 28, pp. 4535-4564, December, 1999. doi: 10.1016/S1352-2310(99)00272-1

[10] B. F. Yu, Z. B. Hu, M. Liu, H. L. Yang, Q. X. Kong, and Y. H. Liu, "Review of research on air-conditioning systems and indoor air quality control for human health", Int. J. Refrig., vol. 32, n^{o}1, pp. 3-20, January 2009, doi: 10.1016/j.ijrefrig.2008.05.004.

[11] D. R. Miller, S. A. Akbar, and P. A. Morris, "Nanoscale metal oxide-based heterojunctions for gas sensing: A review", Sens. Actuators B Chem., vol. 204, pp. 250‑272, December 2014, doi: 10.1016/j.snb.2014.07.074.

[12] J. Zhang, X. Liu, G. Neri, and N. Pinna, "Nanostructured Materials for Room-Temperature Gas Sensors", Adv. Mater., vol. 28, pp. 795‑831, February 2016, doi: 10.1002/adma.201503825.

[13] T. Iwata, K. Matsuda, K. Takahashi, and K.Sawada, "CO_2 Sensing Characteristics of a La_2O_3/SnO_2 Stacked Structure with Micromachined Hotplates", Sensors, vol. 17, n^{o}9, pp. 2156, September 2017, doi: 10.3390/s17092156.

[14] Y. Xiong, Q. Xue, C. Ling, W. Lu, D. Ding, L. Zhu, and X. Li " Effective CO_2 detection based on LaOCl-doped SnO_2 nanofibers: Insight into the role of oxygen in carrier gas", Sens. Actuators B Chem., vol. 241, pp. 725-734, March 2017, doi: 10.1016/j.snb.2016.10.143.

[15] T. Ishihara, "Application of Mixed Oxide Capacitor to the Selective Carbon Dioxide Sensor", J. Electrochem. Soc., vol. 138, n^o 1, pp. 173-176, 1991, doi: 10.1149/1.2085530.

[16] T. Ishihara, K. Kometani, Y. Nishi, and Y. Takita, "Improved sensitivity of CuO-$BaTiO_3$ capacitive-type CO_2 sensor by additives", Sens. Actuators B Chem., vol. 28, n^{o}1, pp. 49-54, July 1995, doi: 10.1016/0925-4005(94)01539-T.

[17] M.-S. Lee and J.-U. Meyer, "A new process for fabricating CO_2-sensing layers based on $BaTiO_3$ and additives", Sens. Actuators B Chem., vol. 68, n^o 1‑3, pp. 293-299, August 2000, doi: 10.1016/S0925-4005(00)00447-0.

[18] P. Keller, H. Ferkel, K. Zweiacker, J. Naser, J.-U. Meyer, and W. Riehemann, "The application of nanocrystalline $BaTiO_3$-composite films as CO_2-sensing layers", Sens. Actuators B Chem., vol. 57, n^{o}1‑3, pp. 39-46, September 1999, doi: 10.1016/S0925-4005(99)00151-3.

[19] J. Herrán, N. Pérez, E. Castaño, A. Prim, E. Pellicer, T. Andreu, F. Peiró, A. Cornet, and J.R. Morante, "On the structural characterization of $BaTiO_3$–CuO as CO_2 sensing material", Sens. Actuators B Chem., vol. 133, n^o 1, pp. 315-320, July 2008, doi: 10.1016/j.snb.2008.02.052.

[20] J. Herrán, G. G. Mandayo, and E. Castaño, "Physical behaviour of $BaTiO_3$–CuO thin-film under carbon dioxide atmospheres", Sens. Actuators B Chem., vol. 127, n^o 2, pp. 370-375, November 2007, doi: 10.1016/j.snb.2007.04.035.

[21] A. M. El-Sayet, F. M. Ismail, and S. M. Yakout, "Electrical Conductivity and Sensitive Characteristics of Ag-Added $BaTiO_3$-CuO Mixed Oxide for CO_2 Gas Sensing", J. Mater. Sci. Technol., vol. 27, n^o 1, pp. 35-40, January 2011, doi: 10.1016/S1005-0302(11)60022-4.

[22] S. B. Rudraswamy and N. Bhat, "Optimization of RF Sputtered Ag-Doped $BaTiO_3$-CuO Mixed Oxide Thin Film as Carbon Dioxide Sensor for Environmental Pollution Monitoring Application", IEEE Sens. J., vol. 16, n^o 13, pp. 5145-5151, July 2016, doi: 10.1109/JSEN.2016.2567220.

[23] S. Joshi, S. J. Ippolito, S. Periasamy, Y. M. Sabri, and M. V. Sunkara, "Efficient Heterostructures of Ag@CuO/$BaTiO_3$ for Low-Temperature CO_2 Gas Detection: Assessing the Role of Nanointerfaces during Sensing by Operando DRIFTS Technique", ACS Appl. Mater. Interfaces, vol. 9, n^o 32, pp. 27014-27026, August 2017, doi: 10.1021/acsami.7b07051.

[24] S. B. Rudraswamy, P. K. Basu, and N. Bhat, "$BaTiO_3$ based Carbon-dioxide gas sensor", in 2012 International Conference on Emerging Electronics, Mumbai, India, December 2012, pp. 1-4, doi: 10.1109/ICEmElec.2012.6636269.

[25] B. Liao, Q. Wei, K. Wang, and Y. Liu, "Study on CuO–$BaTiO_3$ semiconductor CO_2 sensor", Sens. Actuators B Chem., vol. 80, pp. 208-214, December 2001, doi: 10.1016/S0925-4005(01)00892-9.

[26] B. Ostrick, M. Fleischer, H. Meixner, and C.-D. Kohl, "Investigation of the reaction mechanisms in work function type sensors at room temperature by studies of the cross-sensitivity to oxygen and water: the carbonate–carbon dioxide system", Sens. Actuators B Chem., vol. 68, n^o 1‑3, pp. 197-202, August 2000, doi: 10.1016/S0925-4005(00)00429-9.

[27] M.-I. Baraton, L. Merhari, P. Keller, K. Zweiacker, and J.-U. Meyer, "Novel Electronic Conductance CO_2 Sensors Based on Nanocrystalline Semiconductors", MRS Proc., vol. 536, pp. 341, 1998, doi: 10.1557/PROC-536-341.