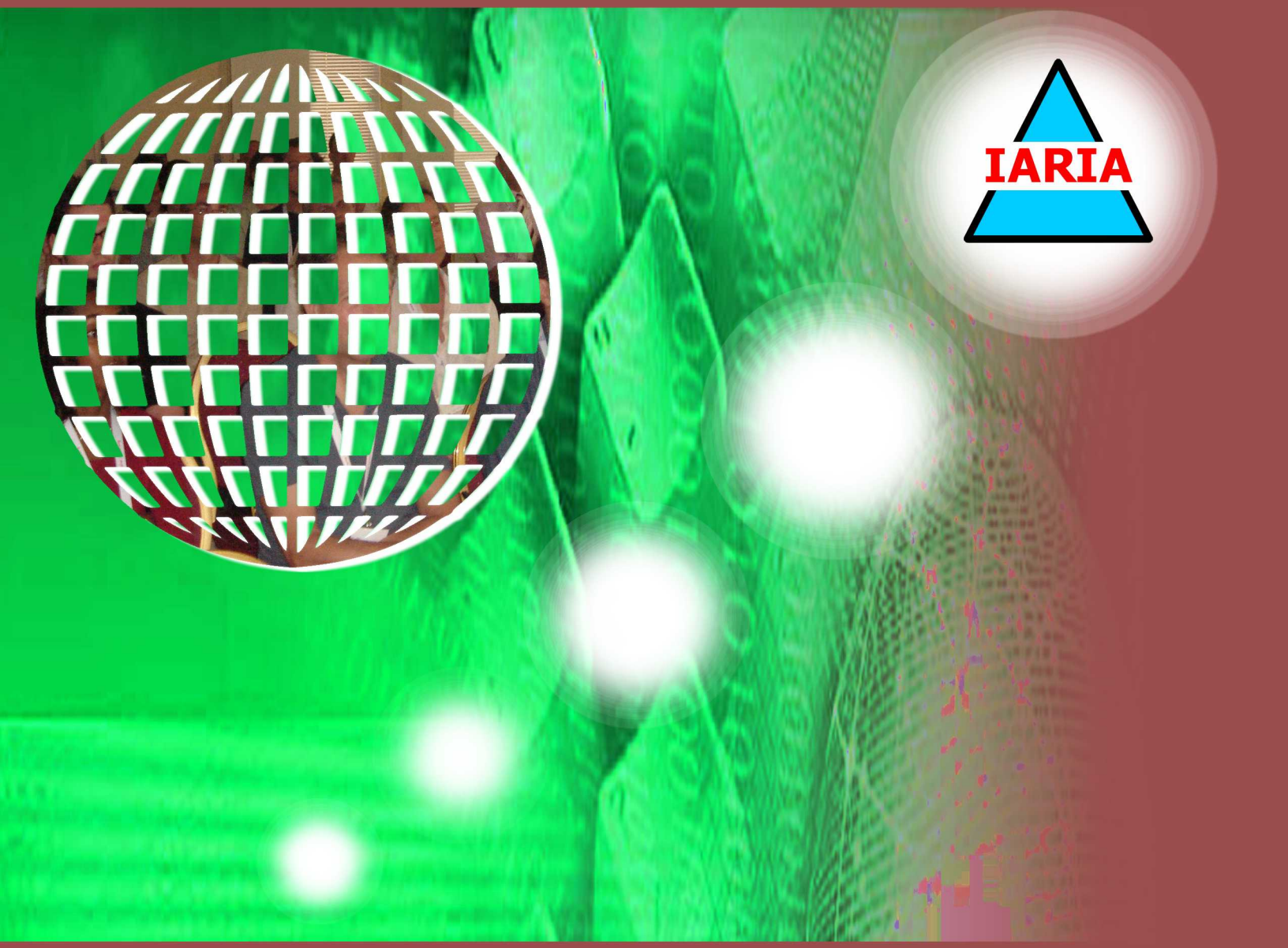
International Journal on Advances in Life Sciences



2010 vol. 2 nr. 3&4

The *International Journal on Advances in Life Sciences* is published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 2, Number 3 & 4, 2010

- Steve Wheeler, University of Plymouth, UK

Advanced Knowledge Representation and Processing

- Freimut Bodendorf, University of Erlangen-Nuernberg Germany
- Borka Jerman-Blazic, Jozef Stefan Institute, Slovenia
- Andrew Kusiak, The University of Iowa, USA
- Selmin Nurcan, University Paris 1 Pantheon Sorbonne, France
- Jeff Riley, Hewlett-Packard Australia, Australia
- Lars Taxen, Linkoping University - Norrkoping, Sweden

International Journal on Advances in Life Sciences
Volume 2, Numbers 3 & 4, 2010

CONTENTS

A Micro-Biomanipulation Training System based on Mixed-Reality

Leonardo S. Mattos, Darwin G. Caldwell
Advanced Robotics Department
Italian Institute of Technology (IIT)
Genoa, Italy
{leonardo.demattos, darwin.caldwell}@iit.it

Abstract— **Within neuroscience, micromanipulation and microinjection of cells (blastocysts and neurons) are essential and highly skilled tasks that can require years of training. These tasks are traditionally performed via direct manual control of the biomanipulation equipment while looking through a microscope. Yet, even after extensive training, yield can be low (40 – 70%). This paper presents a mixed-reality system for the training of operators (biologists/neuroscientists) on a new fully teleoperated biomanipulation system with reduced training requirements and much higher yields. Two mixed-reality training scenarios were designed, implemented and tested for this purpose: A "move-and-inject" task focused on precise positioning training; and a trajectory following scenario intended to develop precise motion control skills in new operators. Preliminary experiments performed with 20 totally novice operators demonstrate that this new training system is effective in terms of the initial development of control skills for real teleoperated biomanipulations. Experimental metrics demonstrate an exponential learning curve for these novice operators, who achieve good performance values after only two practice runs on the system. In addition, this training is shown to be safe and inexpensive since no real cells, biochemical products, or several pipettes are needed for this initial training phase.**

Keywords-mixed-reality; teleoperation; biomanipulation; micromanipulation

I. INTRODUCTION

Biomanipulation, in the context of this work, involves the transportation, orientation and injection of microscopic biological structures such as single cells and early embryos. These operations are normally performed under high-magnification microscopes using glass pipettes with very fine tips (2-50μm), which are attached to micromanipulators and microinjectors. Micromanipulators are mechanical or electro-mechanical devices that scale down the operator's motions to enable precise control of tools in the micrometer range. Microinjectors are devices used for precise control of fluid motion inside the pipettes.

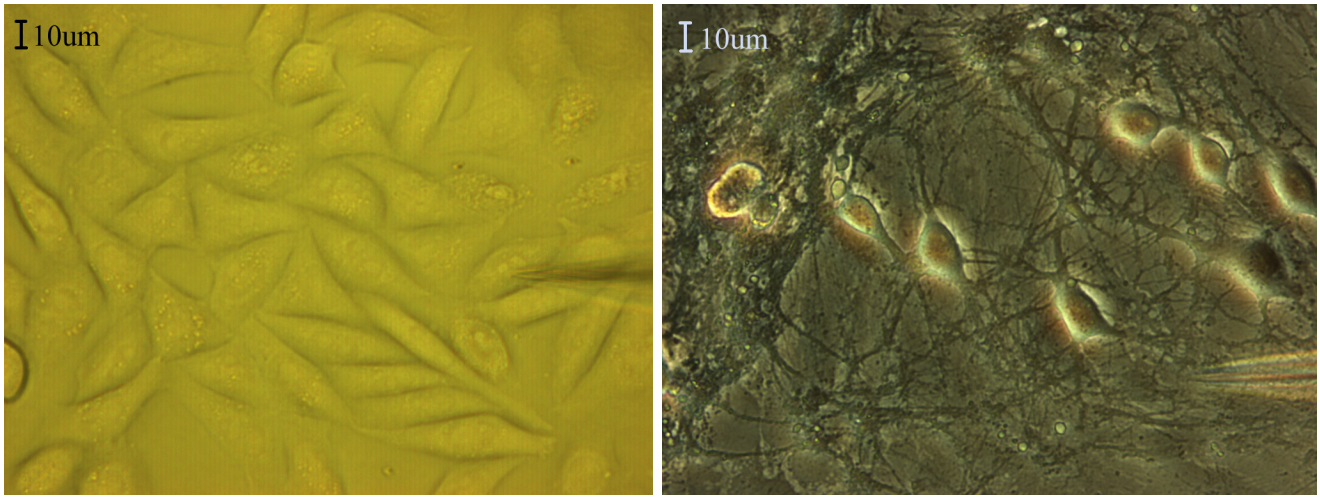
Modern biomanipulation devices are motorized and capable of high-resolution control. For example, commonly used Eppendorf equipment, such as the TransferMan NK2 micromanipulator and the FemtoJet microinjector, offer motion resolution down to 40ηm and the capability of injecting volumes down to the femtolitre into cells. However, under direct operator control these high-resolution motion capabilities are difficult to achieve and do not easily translate into accurate control and successful manipulations.

As a result of these operational difficulties the training period required to achieve proficiency in biomanipulations is normally high, reaching up to one year for operations such as embryo microinjection [2]. In addition, the training process is expensive, involving not only the costs of the underperforming operator, but also expenses associated with wasted materials, samples preparation, cell culture, etc. Furthermore, even after extensive training, the success rates of biomanipulations are found to be less than ideal (40 - 70% for embryo microinjections [3]), pointing to problems related to the user interface and ergonomic factors of the microscope/micromanipulation setup [4]. Typical issues include: high susceptibility to human errors, such as unintentional erroneous motions; and the tiring working conditions, where operators spend hours looking through microscopes while simultaneously controlling micromanipulators and microinjectors.

Improvements to the biomanipulation setup and to its control interface can be achieved using teleoperation techniques whereby the operator controls the system from a computer station, looking at the live video captured from the microscope and displayed on the computer screen. The control of the micromanipulators can be accomplished through the computer keyboard [5]; game joysticks [6]; or even through haptic devices [7][8].

A teleoperated system has the potential to greatly improve biomanipulations by offering supervised control of the micromanipulator motions, which can filter hand tremors and even block erroneous motions [9][10]. In addition, a single teleoperated system can offer different control modalities for the micromanipulators, including position [11], velocity [6] and force control [12], which can be selected according to the specific biomanipulation task or user preference. Furthermore, the speed and precision of the micromanipulator can be dynamically adjusted in a teleoperated system, both automatically [13][14] or manually, further improving manual operations.

Additional benefits of a teleoperated biomanipulation system include the automatic execution of motions that are virtually impossible under direct manual control of the micromanipulators. Examples are simultaneous motions in three dimensions, e.g. fast and precise motions such as "stabbing" movements used to penetrate some cells; and slow linear motions often desired for retracting the pipette from injected cells along the entry path. Another obvious



(a) (b)

Fig.1. (a) Microinjection of CHO-K1 cells under bright-field imaging. (b) Microinjection of neurons using phase contrast imaging. Scale bars on the pictures read 10µm. The diminutive size of these adherent cells require the use of high-magnification objectives, complicating precise pipette positioning over the target cells due to the shallow focal depth of these lenses.

benefit is the creation of a more ergonomic and intuitive work environment by careful arrangement of the display screen and the interfaces to enhance hand-eye collocation and co-ordination. This is less tiring, since the operator can sit comfortably in front of a computer display. In fact, this is a familiar environment to most people, including new operators that should learn biomanipulation tasks, making us hypothesize that teleoperation can enable faster learning at least in part due to this simple reason.

Mixed-reality systems have been developed and used in teleoperations to achieve a diverse range of goals. For instance, they have been used to increase the safety of operations through the creation of virtual barriers in the operating field [10][15]; and also to assist in cell microinjection tasks by providing a preferred direction of motion [3]. Mixed reality has also been used in predictive displays, assisting the planning and execution of robot motions [16]; and in nanomanipulations to provide "haptic" sense in intangible tasks [17]. They have also been used in many training systems for tasks varying from aircraft piloting [18] to minimally invasive surgery [19].

The use of an assistive mixed-reality system can also improve the training and the yield of biomanipulations. For example, our teleoperated system [6] allows the definition of operating zones directly on the live video display, which are used to dynamically adjust the speed and precision of the micromanipulator during operation. This is useful to prevent errors and contributes to increased operation yields because the micromanipulator precision can be automatically increased when the biomanipulation pipette is near a target cell or a danger area.

This paper focuses on the training benefits of a mixed-reality biomanipulation system. Our goal is to show that initial operator training can be done "off-line," without the need for real cells, biochemical products, or the numerous pipettes that are needed when learning biomanipulations on the traditional manually controlled systems. To this end, a fully teleoperated biomanipulation system [6] was used as the real setting for operator training, and target cells were replaced by virtual targets and virtual obstacles. Training was performed in two different mixed-reality scenarios: A move-and-inject task focused on training precise pipette positioning; and a trajectory following scenario, intended to train operators on precise control of the pipette motions. Evaluation of training metrics from 20 totally novice operators are presented here, showing that user learning improved exponentially and demonstrating the great value of this mixed-reality training system.

II. BIOMANIPULATION TASKS

Within the broad field of biomanipulation and particularly microinjection, which is the primary focus of this research, two tasks are of special interest due to their importance and frequency in neuroscience research: adherent cell microinjections and blastocyst microinjections.

Adherent cell microinjection is a delicate and complicated operation that involves the manipulation of cells with dimensions down to 10µm. Figure 1(a) shows adherent cells commonly used in biological and medical

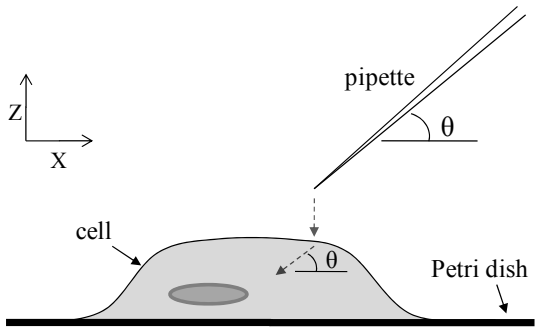
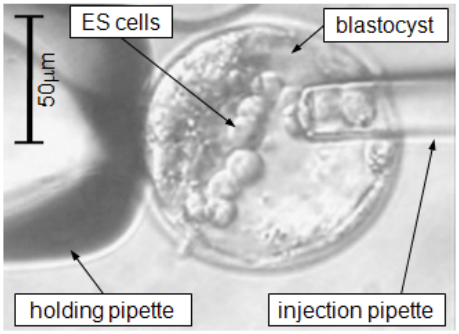


Fig. 2. Synchronized motions on the X- and Z-axis are necessary to create a linear microinjection action when the pipette is positioned on an angle.



BEGIN ***blastocyst microinjection***
CAPTURE blastocyst
ADJUST blastocyst orientation
IMMOBILIZE blastocyst
INSERT needle into the blastocoel cavity
DELIVER around 15 embryonic stem cells
RETRACT needle from blastocyst
RELEASE blastocyst
END

Fig. 3. Blastocyst microinjection and high-level algorithm describing this delicate operation. Coordinated and meticulous control of two micromanipulators and two microinjectors is required in this case, making this operation difficult to master and characterizing its results by low success rates.

research, the CHO-K1 cells [20]. Other examples are neurons, as in Fig.1(b), and endothelial cells. Since these cells adhere to the bottom of the Petri dish, only one pipette/micromanipulator is required to perform the microinjections. However, the pipette needs to be lifted when moving from one cell to another to avoid collisions with other cells and pipette tip breakage. Then, the pipette should be precisely positioned on the upper surface of the next target cell. This operation is complicated by the fact that the diminutive cells require the use of microscope objectives with high magnification factors (40 - 100x), which have a shallow focal depth. This means the pipette tip goes completely out of focus when lifted, making precise positioning difficult. In addition, the injection motion involves simultaneous and coordinated motion in two axes when 3-axis micromanipulators are used, e.g., injection at angle from the top involves simultaneous motions in X and Z (see Fig. 2).

Blastocyst microinjections can be considered as a form of suspension cell microinjection, even though the blastocysts (early embryos) are composed of many cells. In this case, the (mouse) blastocyst moves freely on the bottom of the Petri dish, so injection normally requires the use of two moveable pipettes: one to hold the blastocyst; and the second to perform the actual injection. Figure 3 shows this operation, from which it can be seem that mice blastocysts, measuring around 100µm in diameter, are much larger than the adherent cells mentioned above. Nonetheless, blastocyst microinjections are equally delicate and difficult since even a small error can kill the embryo. One complication here comes from the fact that this task requires coordinated control of two micromanipulators; and another from the need to also control two microinjectors. These complications contribute to increasing the training time required to attain proficiency on this operation, which can typically take up to one year.

As a result of the long training periods and susceptibility to small errors, our research aims at increasing the consistency and efficiency rates attained in manually controlled operations through assisted teleoperation. Consequently, we have developed a teleoperated system that allows easy and precise control of the entire biomanipulation setup from a user-friendly interface. Nonetheless, efficient operation of this new system also requires some training to attain optimum performance, which motivated the development of the mixed-reality training system described in this paper. The next sections present the system created, the training procedure, and initial training experiments performed with completely novice operators.

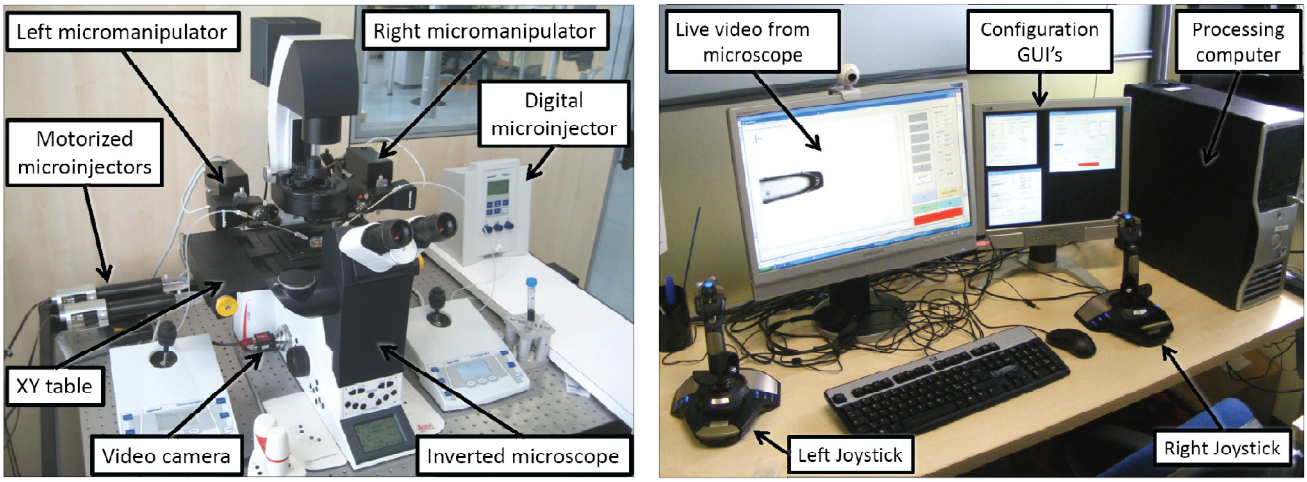
III. BIOMANIPULATION SYSTEM CONFIGURATION

The teleoperated biomanipulation system used in this research was created by the integration of high-end commercial equipment commonly found in neuroscience research laboratories. This was done in favor of: 1) creating a flexible system applicable to a large range of biomanipulation applications; 2) minimizing extra investment from the laboratories that already possess biomanipulation equipment; and 3) increasing the system's acceptance by the biology/neuroscience community, which is already familiar with and trusts the equipment used.

Equipment selection and configuration was performed in collaboration with neuroscience researchers. Based on the mix of their research needs with engineering specifications for automation, the developed system included:

- one Leica DMI6000B inverted microscope
- two Eppendorf TransferMan NK2 motorized micromanipulators
- one Eppendorf Femtojet microinjector
- one Marzhauser SCAN IM 120x100 motorized scanning microscope stage
- two Eppendorf CellTram Vario microinjectors incorporating custom computer-controlled driving systems
- a desktop computer with an Intel Core2 Quad 2.83 GHz CPU, WindowsXP, and 3GB RAM
- an AVT Guppy F-080C firewire camera
- two Saitek Cyborg Evo Force joysticks

These devices are shown in Fig. 4, which presents the two workstations that constitute the developed teleoperated biomanipulation system: The microscope station and the control station.



(a) (b)

Fig. 4. Teleoperated biomanipulation system configuration: (a) the microscope station; (b) the computer/control station.

The selected devices are appropriate for many different biomanipulation procedures, including the microinjection of blastocysts and adherent cells. These operations can be precisely performed using the selected micromanipulators (which have three translation axes with resolution of 40ηm and a maximum speed of 7.5 mm/s) and the FemtoJet microinjector, which can deliver volumes down to the femtolitre.

IV. System Control and Teleoperation

Control and teleoperation of the devices were achieved through standard 2-way communication interfaces, including RS-232C and CAN. This was possible because all of the selected devices supported external control through a serial communications port. Therefore, the type of connection between the controlling computer and a system device was dictated by the device's supported interface.

All system devices were integrated by software, and could be simultaneously controlled from a graphical user interface (GUI) running on the desktop computer (see Fig. 5). User commands were received via the joysticks, computer mouse or computer keyboard. These commands were processed and then forwarded to the appropriate biomanipulation equipment.

Feedback from the micro-world was obtained through the video camera, which provided live video feed from the microscope's field of view.

The virtual features used during this research were created using graphics overlaid on the live video stream. The interaction of these virtual features with real system components created the mixed-reality biomanipulation environment. These interactions were enabled by mapping the micromanipulator coordinates to the image space, as described in [21] and summarized below. Using this mapping the system was able to compute the image coordinates of the tool (pipette) without the need for image-based localization software. This created a robust and fast system capable of generating an operating environment influenced by both real and virtual objects. Figure 5 shows some of the virtual features that could be created, including;

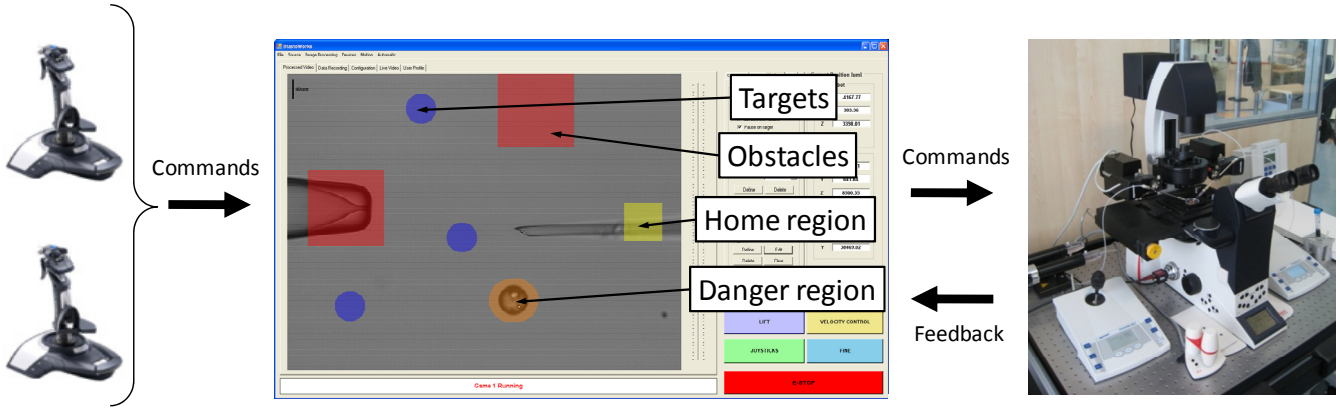


Fig. 5.The mixed-reality biomanipulation system setup and examples of virtual features defined on the operating environment.

targets, obstacles and a danger region. All regions were customizable in terms of maximum speed allowed and repelling force. These virtual components were the basis of the two mixed-reality training scenarios described in section V.

A. System Calibration

The calibration procedure used to compute the mapping between the video and robot coordinate frames assumed zero image distortion by the system optics and perfect parallelism between the camera plane and the robot's X-Y plane. Considering these two simplifying assumptions, the calibration procedure consisted of finding the translation, rotation and scaling factors required to map points between the two frames. This was realized by requesting the user to click on the tool tip seen on the live video at five different robot locations. These operations defined $\{P_0, P_1, P_2, P_3, P_4\}$, five reference points described with coordinates ${}^{V}P_n$ in the video frame and ${}^{R}P_n$ in the robot frame. The first two points were initially used to compute the rotation, θ, between the coordinate frames according to the method described in Fig. 6; and later to compute the frame transformation described by:

$$ {}^{V}P = S \bullet {}^{V}_{R}Rot \bullet \left({}^{R}P - {}^{R}P_0 \right) + {}^{V}P_0 \quad (1) $$

where P is a point of interest; ${}^{V}_{R}Rot$ is the rotation matrix from the robot to the video coordinate frame; and S is the scaling matrix. These were defined as:

$$ {}^{V}_{R}Rot = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (2) $$

$$ S = \begin{bmatrix} S_X & 0 \\ 0 & S_Y \end{bmatrix} \quad (3) $$

The scaling factors S_X and S_Y were measured in pixels/micrometers. They were computed using Eq. 1 and the acquired points P_0 and P_1.

After this initial mapping estimation, the entire set of reference points was used to improve the mapping parameters. This was achieved by minimization of the error function defined by (4), which represents the RMS error between the real and the computed tool positions in robot coordinates.

$$ \epsilon = \sqrt{\frac{\sum_{i=0}^{n}\left\{\left({}^{R}\hat{X}_i - {}^{R}X_i\right)^2 + \left({}^{R}\hat{Y}_i - {}^{R}Y_i\right)^2\right\}}{n}} \quad (4) $$

In (4), $\left({}^{R}X_i , {}^{R}Y_i\right)$ represents the i^{th} actual x-y robot position, $\left({}^{R}\hat{X}_i , {}^{R}\hat{Y}_i\right)$ represents the i^{th} computed x-y robot position, and n is the number of calibration points used for the computations.

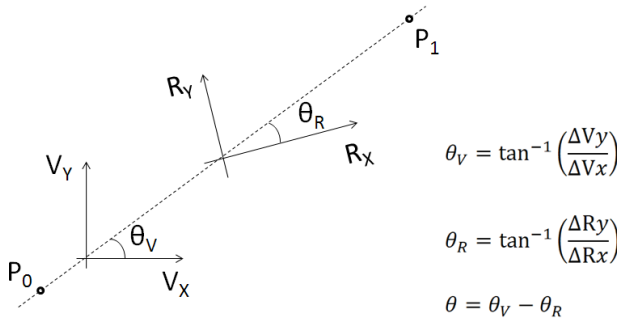


Fig. 6: Method used to compute the rotation between the robot and the video coordinate frames. P_0 and P_1 were points with known coordinates on both frames.

V. TRAINING SCENARIOS

The main goal of the developed training scenarios was familiarize new users with the teleoperated system environment, introducing them to the system control and joystick functions. Training using mixed-reality is a safe, fast and inexpensive way of developing the necessary control skills with new operators. Therefore, two tasks were developed to provide the basic training that could guarantee high levels of performance and safe operations even on the very first real biomanipulations.

For evaluation purposes, both tasks always start at the same point, i.e., with the pipette tip inside a fixed virtual home region (see Fig. 5). In addition, the last task of a training session has the exact same configuration (in terms of target locations or trajectory to be followed) as the first test attempted, enabling a fair comparison between initial and final operator performances.

A. Move-and-inject scenario

This task was created to focus operator training on precise teleoperated positioning of the pipette tip. In this case, several virtual targets are randomly overlaid on the live video captured from the microscope's field of view, as shown in Fig. 5. The task of the operator is to move the pipette's tip to each of these targets. When the pipette is over a target, the operator should "inject" it by pressing the joystick's trigger button. This action corresponds to an automatic injection motion when manipulating real cells, which is customizable in terms of distance, direction and speed, and is triggered by the same joystick button. When a target is successfully injected, it disappears from the screen. The operator's goal is to eliminate all targets displayed. Completion of this goal finalizes the task.

Different levels of difficulty can be set in this scenario by changing: 1) the number of targets; 2) the size of the targets; 3) the number of obstacles; 4) the size of the obstacles; and 5) the maximum time allowed to complete the task. Making the targets smaller increases task difficulty because it requires better precision in pipette positioning. The presence of obstacles also increases game difficulty because this means the user has to learn to control the pipette's trajectory, not simply go straight to the targets.

Experimental data collected during this task includes the pipette tip position in image coordinates, recorded at 25Hz; the number of collisions with virtual obstacles; and the total duration of the task, i.e., the amount of time spent to "inject" all targets. Saving the pipette tip positions enables offline analysis of the user's skills and strategies. Here, these data were used to measure the total distance travelled by the pipette during the test, which was then normalized by the absolute minimum travel distance computed using a version of the travelling salesman algorithm. This normalized distance was used as a performance metric. In addition, the average pipette speed during the experiment was also computed and used as a performance metric.

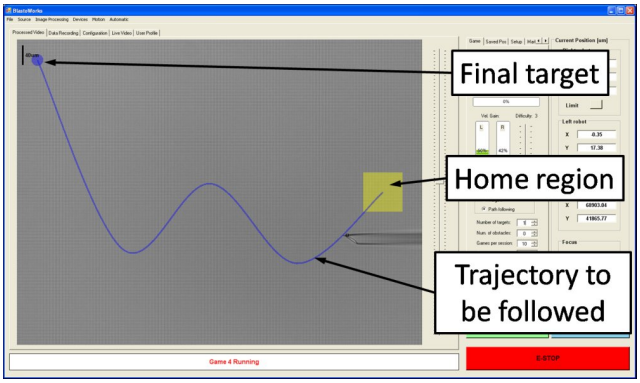


Fig. 7. The trajectory following game scenario.

B. Trajectory following scenario

The goal of this task was to train new users on the dynamic control of the micromanipulation pipette motions, focusing on speed and direction control skills. Here, a desired trajectory was randomly defined, but always started in the centre of the home region and finished on a target located at the opposite side of the video panel. The operator's task consisted of guiding the pipette tip from the

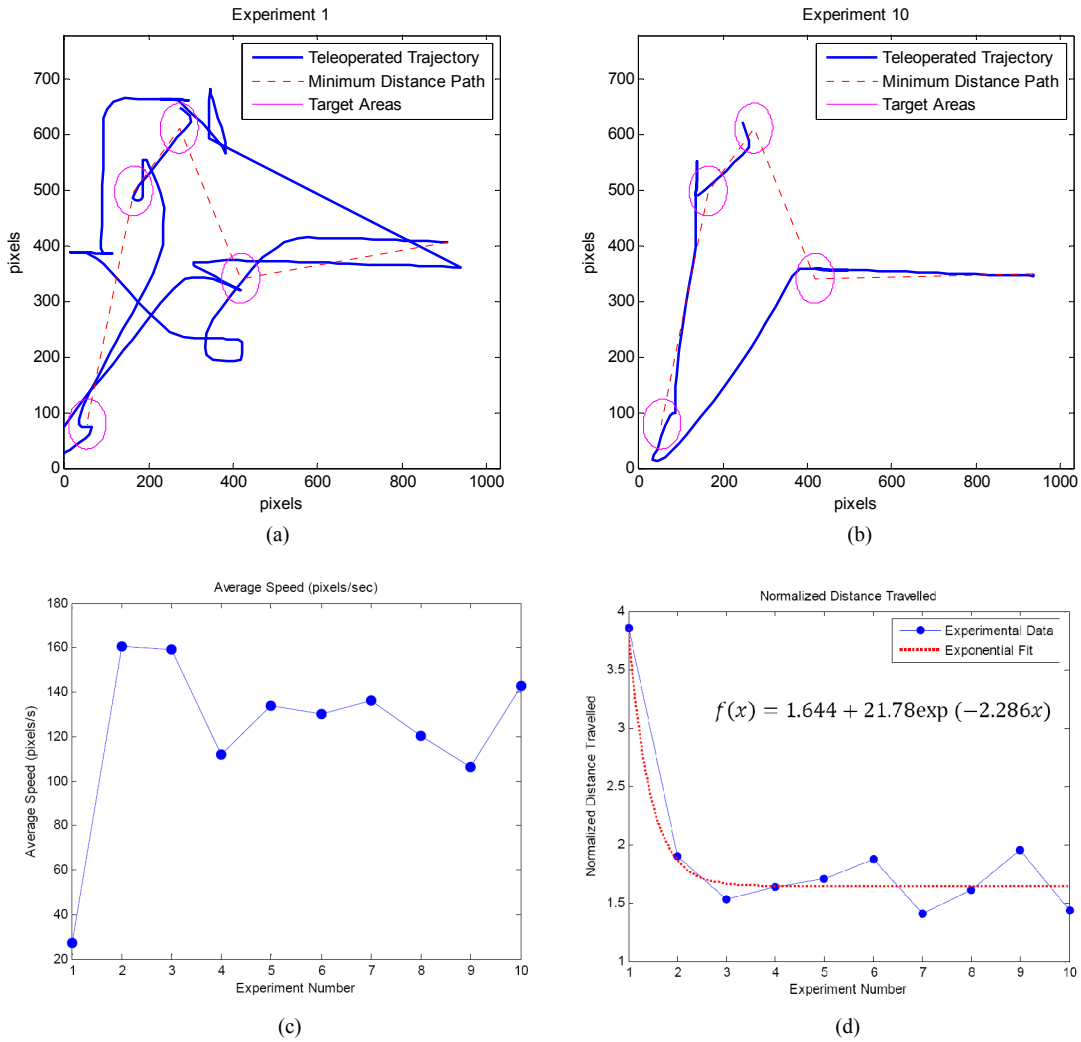


Fig. 8. Examples of game data and metrics collected for the move-and-shoot game played by a novice operator: (a) Data from the first training session game; (b) Data from the last training session game; (c) Average pipette speed on each game played; (d) Normalized distance travelled on each game played.

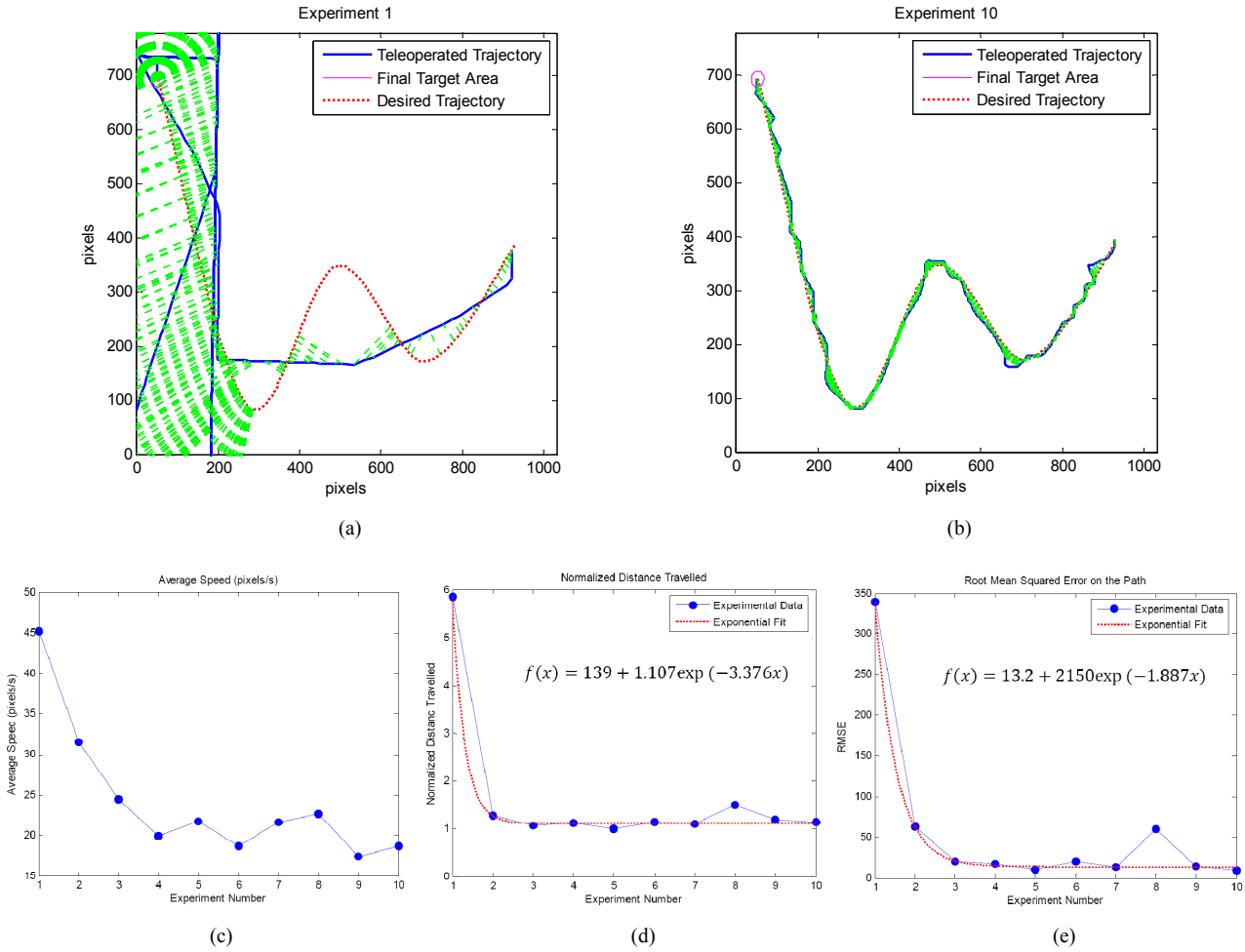


Fig. 9. Examples of task data and metrics collected for the trajectory following scenario played by a novice operator: (a) Data from the first training session, with green lines showing the smallest distance from a sampled pipette coordinate to the desired trajectory; (b) Data from the last training session; (c) Average pipette speed on each task played; (d) Normalized distance travelled; (e) RMSE of the trajectory following task.

home region to the final target following the given trajectory. A typical task is shown in Fig. 7.

As in the previous scenario, the difficulty level could also be adjusted. This was achieved by changing the number of waypoints selected for the trajectory definition. Increasing the number of waypoints creates more twists and turns on the desired trajectory, requiring better motion and speed control to achieve a good performance. The trajectory was defined using spline interpolation, i.e., as a smooth curve passing through all given waypoints. Randomness came from the fact that the y-coordinate of the waypoints were defined randomly (but not the x-coordinates – these were regularly spaced between the home region and the target).

Evaluation of the operator performance was based on the root-mean-squared-error (RMSE) between the trajectory followed and the desired trajectory. This error was computed offline, based on the recorded pipette tip coordinates. The error from a sampled pipette coordinate to the desired trajectory was assumed to be the smallest distance between those two entities. This measure is represented by the green lines in Fig. 9, which displays experimental data from one of the novice operators.

Another two measures obtained from this task were the total distance travelled by the pipette and the amount of time required to complete the given task. These were used to generate two metrics: the average pipette speed; and the normalized distance travelled, which was computed using the length of the desired trajectory as the normalizing factor.

VI. TRAINING EXPERIMENTS

The two mixed-reality training scenarios were used to train 20 totally novice operators on the control of the teleoperated biomanipulation system. All of these operators had no prior experience with biomanipulations or any other type of micromanipulation, and each of them went through only one training session in the teleoperated system. Here, a training session consisted of attempting only one of the scenarios 10 times. Therefore, the group of novice operators

was divided in two sets for the experiments: One half performed the move-and-shoot scenario and the other attempted the trajectory following scenario.

The difficulty level of the tasks was kept constant during the training sessions, and was the same for all users. For the move-and-shoot task, the number of targets selected was four; the diameter of the targets was set to 90 pixels; no obstacle was present; and there was no time limit to finish the task. For the trajectory following case, the desired path was defined from four random waypoints plus the fixed initial point at the centre of the home region. Once again there was no time limit to finish the task.

Examples of "move-and-inject" tests and of the data collected during a training session are shown in Fig. 8. Both the first and the last test completed by an operator are presented, from which a great performance improvement can be readily seen. The metrics obtained during this training session show an exponential learning curve, which was typical for operators that classified themselves as non video game players. On average, operators reached a performance level within 10% of their final performance by the third test. An interesting observation was the good performance of operators that classified themselves as gamers. In these cases, little improvement was noticed during the training sessions because these operators performed well from their very first trial.

An example of a trajectory following training session is presented in Fig. 9. In this case a great performance improvement is also clearly seen when comparing data from the first and last games played by the operator. Additionally, the experimental metrics show, once again, an exponential learning curve, which was typical for the non-gamer operators. This learning trend allows us to conclude that the teleoperated biomanipulation system is user-friendly and easy to operate.

The overall average of the trajectory following RMSE computed for all operators decreased from an initial 66.9 pixels to a final 20.3 pixels, demonstrating an error reduction of almost 70% after only one training session. In addition, the overall change in normalized distance travelled between the first and the last scenario of the training sessions was -25.2% for the move-and-inject task, and -21.5% for the trajectory following task, indicating good improvements in pipette motion control for both groups of operators. Another interesting observation from overall average data was the speedup measured for the move-and-inject scenario. On average, the operators are 66.3% faster in precise pipette positioning after undergoing the move-and-inject training session. A much smaller speedup was found for the trajectory following scenario, only 5.5%, which can be explained by the fact that operators learned to keep the speed low to better control the pipette trajectory. These data are summarized in Table I.

TABLE I. OVERALL AVERAGE OF GAME METRIC CHANGES[A]

Metric	Percentage Change	
	Move-and-shoot game	***Trajectory following game***
Average Speed	+66.3%	+5.5%
Normalized Distance Travelled	-25.2%	-21.5%
Trajectory Following RMSE	—	-48.9%

a. computed from data from the first and last games played by each operator.

VII. CONCLUSION AND FUTURE WORK

A mixed-reality training system for teleoperated biomanipulations has been developed and tested during this research. The training platform consisted of a previously developed fully teleoperated biomanipulation system, which was augmented by a new mixed-reality interface developed for operator training. Here, the biomanipulation system provided the real setting for training, while virtual targets and virtual obstacles replaced the real cells to be manipulated. Setup time for training in this system was short, only 3 to 5 minutes, and the pipette was practically impossible to break because it was positioned far away from any physical obstacle.

Two mixed-reality training scenes were designed, implemented and tested during this research: One focused on precise positioning training using a "move-and-inject" task; and the other aimed to develop precise motion control skills in new operators, based on trajectory following tasks. Results from preliminary experiments with 20 totally novice operators demonstrated that this training system was effective in terms of initial development of the necessary control skills for real teleoperated biomanipulations. Training here was shown to be fast, safe, and inexpensive since no real cells, biochemical products, or pipettes were needed for this initial phase.

The experiments demonstrated that learning on this new system was exponential, enabling operators to reach, on average, a performance level within 10% of their final performance by the third training run. In addition, all operators were able to achieve precise positioning at an average rate greater than 8.18 targets/min, and trajectory following with RMSE less than 27.9 pixels after only 10 practice runs. Furthermore, operators that classified themselves as gamers demonstrated this level of performance from their very first trial. These observations not only reiterated that training on the developed mixed-reality system is fast, but also tells us that, as younger generations become more familiar with games, virtual realities, and the use of technologies, the value of a teleoperated system that feels like a computer game tends to increase.

Future training sessions will evaluate the impact of progressively increasing the level of difficulty of the developed scenarios. The goal will be to maintain the operator's initial learning rate for a longer period of time, hopefully taking their final control skills to a more advanced level without increasing the number of practice sessions. In addition, a group of operators will be trained on a mix of both scenarios, and later will be asked to perform real cell

microinjections. This will be the final training system test, which will evaluate the hypothesis that teleoperated skills acquired in the mixed-reality trainer do transfer to real biomanipulations.

REFERENCES

[1] L. S. Mattos and D. G. Caldwell, "A Mixed Reality Training System for Teleoperated Biomanipulations," *Proceedings of International Conferences on Advances in Computer-Human Interactions*, ACHI 2010, February 2010.

[2] D. Kim, S. Yun, B. Kim, "Mechanical force response of single living cells using a microrobotic system," *2004 IEEE International Conference on Robotics and Automation*, New Orleans, USA, April 2004.

[3] A. Kapoor, R. Kumar, R. H. Taylor, "Simple biomanipulation tasks with Steady Hand cooperative manipulator," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*, Lecture Notes in Computer Science, vol. 2878, pp. 141-148, ISBN: 3-540-20462-8, 2003.

[4] N. Stylopoulos and D. Rattner, "Robotics and ergonomics," *Surgical Clinics of North America*, vol. 83(6), pp. 1321-1337, 2003.

[5] S. Ogawa, H. Takahashi, J. Mizuno, N. Kashiwazaki, M. Yamane, E. Narishige, "Personal computer-controlled microsurgery of fertilized eggs and early embryos," *Theriogenology*, vol. 25(2), 1986.

[6] L. Mattos and D.G. Caldwell, "Interface design for microbiomanipulation and teleoperation," *Second International Conference on Advances in Computer-Human Interactions*, ACHI 2009, Mexico, February 2009.

[7] D. Kim, B. Kim, S. Yun, and S. Kwon, "Cellular force measurement for force reflected biomanipulation," *2004 IEEE International Conference on Robotics and Automation*, New Orleans, USA, April 2004.

[8] A. Pillarisetti, M. Pekarev, A. D. Brooks, J. P. Desai, "Evaluating the effect of force feedback in cell injection," *IEEE Transactions on Automation Science and Engineering*, Vol. 4, No. 3, July 2007.

[9] G. H. Ballantyne and F. Moll, "The da Vinci telerobotic surgical system: the virtual operative field and telepresence surgery," *Surgical Clinics of North America*, vol. 83(6), pp. 1293–1304, 2003.

[10] L. B. Rosenberg, "Virtual fixtures: Perceptual tools for telerobotic manipulation," *IEEE Virtual Reality Annual International Symposium*, pp. 76–82, 1993.

[11] O. Tonet, M. Marinelli, G. Megali, A. Sieber, P. Valdastri, A. Menciassi, P. Dario, "Control of a teleoperated nanomanipulator with time delay under direct vision feedback," *2007 IEEE International Conference on Robotics and Automation*, pp. 3514-3519, Rome, Italy, 2007.

[12] X.Y. Liu, K.Y. Kim, Y. Zhang, and Y. Sun, "NanoNewton force sensing and control in microrobotic cell manipulation," *International Journal of Robotics Research*, vol. 28(8), pp. 1065-1076, 2009.

[13] N. Turro, O. Khatib, E. Coste-Maniere, "Haptically Augmented Teleoperation," *2001 IEEE International Conference on Robotics and Automation*, pp. 386- 392, 2001.

[14] N. Garcia-Hernandez and V. Parra-Vega, "Haptic Teleoperated Robotic System for an Effective Obstacle Avoidance," *Second International Conference on Advances in Computer-Human Interactions*, ACHI 2009, Mexico, February 2009.

[15] B.L. Davies, S. Harris, M. Jakopec, J. Cobb, "A novel hands-on robot for knee replacement surgery," *Computer Assisted Orthopaedic Surgery USA* (CAOS USA), A. DiGioia and B. Jaramaz, Eds. Pittsburgh: UPMC Medical Center, pp. 70-74, 1999.

[16] A.K. Bejczy, W.S. Kim, S. Venema, "The phantom robot: predictive displays for teleoperation with time delay," *1990 International Conference on Robotics and Automation*, 1990.

[17] M. Ammi, A. Ferreira, J-G. Fontaine, "Virtualized reality interfaces for telemicromanipulation," *2004 IEEE International Conference on Robotics and Automation*, New Orleans, USA, April 2004.

[18] Finnair Flight Training Center, www.finnairflighttraining.com

[19] G. Lacey, D. Ryan, D. Cassidy, D.Young, "Mixed-reality simulation of minimally invasive surgeries," *IEEE MultiMedia*, vol.14(4), p.76-87, October 2007.

[20] K. Jayapal, K. Wlaschin, M. Yap, W-S Hu, "Recombinant protein therapeutics from CHO cells - 20 years and counting," Chemical Engineering Progress, vol. 103(7), October, 2007.

[21] L. Mattos and D. G. Caldwell, "A fast and precise micropipette positioning system based on continuous camera-robot recalibration and visual servoing," *IEEE Conference on Automation Science and Engineering*, CASE 2009, August 2009.

A Biologically Accurate Simulation of the Locomotion of *Caenorhabditis elegans*

Roger Mailler, Jacob Graves, Nathan Willy, and Trevor Sarratt
Computational Neuroscience and Adaptive Systems Laboratory
University of Tulsa
Tulsa, United States
{*mailler, jacob-graves, nathan-willy, trevor-sarratt*}*@utulsa.edu*

***Abstract*—The nematode *Caenorhabditis elegans* is an important model organism for many areas of biological research including genetics, development, and neurobiology. It is the first organism to have its genome sequenced, complete cell ontogeny determined, and nervous system mapped. With all of the information that is available on this simple organism, *C. elegans* may also become the first organism to be accurately and completely modeled *in silico*. This work takes a first step toward this goal by presenting a biologically accurate, 3-dimensional simulated model of *C. elegans*. This model takes into account many facets of the organism including size, shape, weight distribution, muscle placement, and muscle force. It also explicitly models the environment of the worm to include factors such as contact, friction, inertia, surface tension, and gravity. The model was tuned and validated using video recordings taken of the worm to show that it accurately depicts the physics of undulatory locomotion used to forward and reverse crawl on an agar surface. The main contribution of this article is a new, highly detailed 3D physics model and supporting simulator that accurately reproduces the physics of *C. elegans* locomotion.**

***Keywords*-Simulation, Biology, *Caenorhabditis elegans*, Modeling**

I. Introduction

Nearly 50 years ago, Sydney Brenner introduced *Caenorhabditis elegans* as a model for studying developmental biology and neurology. Because of its simplicity, it has become one of the best understood organisms on the planet being the only one to have its cell lineage, genome, and nervous system completely mapped. However, despite all of the effort that has gone into uncovering the secrets behind "the mind of the worm," we still lack a compelling systems-level understanding for how the neurons and the connections between them generate the surprisingly complex range of behaviors that are observed in this relatively simple organism.

One potential approach to addressing this issue is to use computer simulations that model aspects of the worm's body and nervous system [1]. For example, numerous computer simulations have been created that replicate the locomotion of *C. elegans* [2], [3], [4], [5].

Like all models, these simulations make simplifying assumptions that make them computationally tractable at the expense of accuracy. Each of them, for instance, represents the body as a set of uniformly distributed points in two-dimensional space. This prevents them from replicating the proper weight distribution, and more importantly, the non-uniform placement of the muscles that are used to generate locomotive force in the actual worm. In addition, they also fail to directly simulate the environment, but instead apply constant frictional forces at these discrete points along the body. These simplifying assumptions limit the ability of these simulations to accurately depict the non-uniform friction that results from the worm's contact with the world around it and subsequently the complex neural control that is needed to generate the worm's characteristic sinusoidal pattern of locomotion.

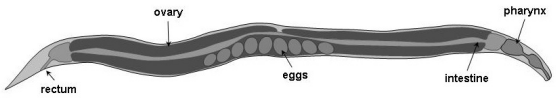


Figure 1. Basic anatomy of an adult hermaphrodite.

Leveraging the tremendous increases in computational power and advances in numeric methods, this work, which is an extended version of the work presented in [1], seeks to rectify these deficiencies by developing a biologically accurate 3D model of the body of *C. elegans* in a virtual environment that mirrors the physical properties of its natural world. This simulator, which has been under development for nearly two years, is built using an open-source 3D game and physics engine. The model accurately depicts the physical properties of the real organism including its non-uniform weight, size, shape, and musculature. In addition, the simulator models the interaction between the worm and its environment to include surface tension, friction, inertia, and gravity.

This paper presents this new model and demonstrates that it faithfully reproduces forward and reverse crawling of *C. elegans* on an agar surface. The model is cross validated using video recordings of worms that were converted to quantitative data by image analysis software. During our validation, we found that, during forward locomotion, the forces generated by the muscles may decrease as the wave propagates from the worm's head to its tail. Although we

found no mention of this in the literature, the placement of the muscles in the worm's body, along with video analysis of the worm's crawling gait seem to support our finding. We also have found that in order to replicate reverse locomotion that the force generated by the muscles needed to be higher, the wavelength shorter, and the wave propagation slower.

The rest of this article is organized as follows. In Section II, a brief introduction to the anatomy of C. elegans is given along with a review of other approaches to modeling this organism. In Section III, we present the underlying simulation technology used in this work as well as provide a detailed description of the physics model. In Section IV, the methods and techniques used to tune and validate the system are described with Section V discussing the results. Finally, in Section VI, we present our conclusions.

II. Background

This section provides the necessary background on the anatomy of *C. elegans* as well as the state-of-the-art in computer simulations of this organism.

A. Caenorhabditis elegans *Anatomy*

Caenorhabditis elegans is a small (1 millimeter in length) nematode that can be found living in the soil of many parts of the world. It lives by feeding on bacteria and is capable of reproducing in about 3 days under the right conditions. *C. elegans* can either be male or hermaphrodite, with males occurring at a low frequency in the population. They reproduce by either self fertilization in the hermaphrodite or by mating a male and hermaphrodite. Hermaphrodites lay about 300 eggs during their approximately 15 day lifespan.

C. elegans (see Figure 1) is a very simple organism, with only 959 cells in the adult hermaphrodite and 1031 in the adult male [6]. Like other members of the nematode (Nematoda) family, the body of *C. elegans* is composed of two concentric tubes separated by a pseudocoelom. The inner tube is in the intestine and the outer tube consists of the hypodermis, muscles, nerves and the gonads. The pseudocoelom is filled with a hydrostatically pressurized fluid that helps maintain the shape of its body.

C. elegans maintains an outer cuticle, which is secreted by the hypodermis. During the lifespan of the worm, it molts its cuticle four times, punctuating the four phases of its life cycle. The cuticle, containing mostly collagen, is tough although not rigid. Adult nematodes have lateral, longitudinal seam cells on the surface of their cuticle that form treads (alae). When on a solid surface, the nematode crawls on one side with a set of treads contacting the surface.

The main body wall muscles of *C. elegans* are arranged in four rows, two dorsal and two ventral. Each row consists of 23 or 24 muscle cells that are arranged in an interleaving pattern [7]. Toward the anterior of the worm, the cells occur in overlapping pairs with less overlap and pairing occurring toward the posterior. The worm moves by propagating waves either forward or backward along its body creating a sinusoidal pattern of locomotion.

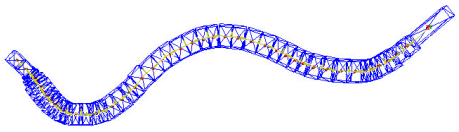


Figure 2. Physics model of the nematode *C. elegans*

C. elegans has a simple nervous system consisting of 302 neurons with about 2000 neuromuscular junctions, 5000 chemical synapses, and 700 gap junctions in the adult hermaphrodite [8]. Most of these cells are located near the pharynx and in the tail. Processes from these neurons form a "nerve ring" that surrounds the pharynx or are part of bundles that run the length of the body. The most noticeable of these bundles are the ventral and dorsal nerve cords.

The nervous system receives input primarily from sensilla located in the head of the worm that are connected to sensory neurons that extend from the nerve ring. The nerve ring sends its output through motor neuron axons that are in the ring itself or located in the ventral or dorsal nerve cords. Most neurons in *C. elegans* have simple structures with one or two processes [9]. Despite their apparent simple structure, neurons in *C. elegans* have a diverse set of voltage-gated, chemically-gated, and mechanically-gated ion channels, use many of the neurotransmitters found in vertebrates, and exhibit a complex mechanism for vesicle production, docking, priming, and release. There is considerable evidence the neurons in *C. elegans* use acetylcholine, GABA, dopamine, serotonin, glutamate, and a set of peptides in communicating with one another [8]. Acetylcholine is used as the primary excitatory neurotransmitter in motor neurons [10]. GABA is used as both an inhibitory and excitatory neurotransmitter [10]. Dopamine appears to influence egg-laying and exists in the male reproductive apparatus [11], [12]. At least ten cells in hermaphrodites seem to signal with serotonin. The strongest influence appears in the pharynx [13], although it influences egg-laying behavior as well [14]. Finally, glutamate seems to be used as both an excitatory and inhibitory neurotransmitter in motor and sensory neurons [15].

Neurons in *C. elegans* appears to contain both voltage-gated potassium and calcium channels [10], but recent advances in electro-physiological techniques have demonstrated that they seem to lack voltage-gated, sodium channels [16]. Because of this, these studies have found that neurons in the organism lack traditional, fast action potentials found in vertebrates. Instead, patch-clamping studies have been able to demonstrate non-linear, graded depolarization as a result of outward potassium flow, which is regulated by inward calcium flow [16].

Even though the nervous system of *C. elegans* has been

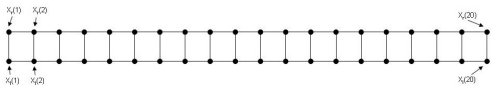


Figure 3. Uniform, two-dimensional model of *C. elegans*

physically mapped [9], the properties, roles, and interdependencies of many of the neurons are still unknown. Much of what is known about the role of individual neurons in *C. elegans* comes from behavioral studies of worms that have a genetic mutation or have undergone laser micro-dissection. As the name implies, laser micro-dissection works by killing an individual cell or group of cells using a laser [17]. Once the cells have been ablated, the behavior of the worm is observed and the role of the neuron identified. For example, using laser ablation, the role of many of the cells in pharyngeal pumping have been determined [18] and several classes of inter-neurons (AIY, AIZ, and RIB) have been found to be involved in controlling locomotion [19], [20]. In addition, the suspected roles of the D-type motor neurons have been uncovered using this technique [21]. More recent work has uncovered the importance of mechano-sensation in determining swimming gait in the worm by ablating the ALM touch receptor neuron [22].

There are a number of ways to measure the behavior of *C. elegans*. The most prominent techniques that are used include measuring changes in frequency, amplitude, or gait during locomotion [23], [3] and using chemotaxis and thermotaxis assays [24].

Even with these techniques, many questions about the function of the nervous system remain unanswered. For example, there are currently three hotly debated theories that attempt to explain how the nervous system generates and propagates waves through the worm's body given its pattern of connectivity. The first theory is that the gap junctions between muscle cells or the motor neurons aid in the propagation of the wave [9], [25]. The second theory is that *C. elegans* undulatory motion could be controlled by a central pattern generator (CPG) like the related nematode Ascaris [26], [27]. The third theory is that stretch receptors, mechanically-gated ion channels, contained within A-type and B-type motor neurons "sense" a wave and then propagate it.

B. C. elegans Simulators

All of the simulators that have, thus far, been created for *C. elegans* are designed to address specific questions about the biology of the worm that cannot be answered using standard biological techniques. The earliest use of a *C. elegans* specific simulator can be found in the work of Niebur and Erdos [2], [27]. Their simulator was designed to investigate the physical mechanics that are used to propel the organism and to determine if stretch receptor control

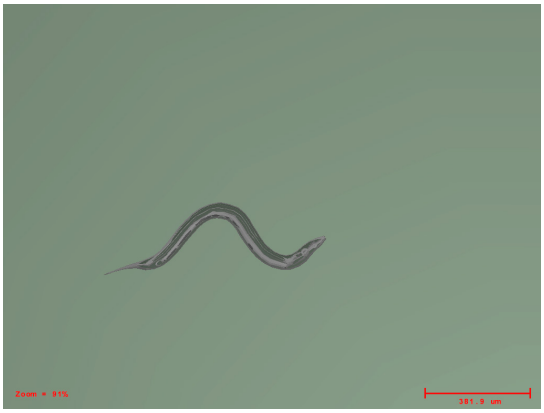


Figure 4. Screenshot of the ALIVE Simulator

could be responsible for generating the worm's characteristic sinusoidal wave.

To show that stretch receptors could create the sinusoidal wave, Niebur and Erdos created a simplified two-dimensional model of the worm that represents the contact between the worm's body and an agar surface (see Figure 3). Their model uses 40 points, 20 on the left and 20 on the right, which are evenly distributed along the length of the body. Each point is connected to the points that are contra-lateral, anterior, and posterior to it by springs that represent the cuticle of the worm. The model is powered by muscles that are connected to the points along the length of the worm. Locomotion occurs when the simulated nervous system causes the muscles to contract. The simulator then updates the position of each of the points based on the forces being created by the muscles, the elasticity of the cuticle, the interior pressure, and the frictional forces of the body's contact with the surface.

Although they did not create an accurate model of the nervous system, Niebur and Erdos were able to demonstrate that sinusoidal waves could be produced by stretch receptors located on the motor neurons that control the body wall muscles of *C. elegans*. Other researchers have extended their model in an effort to further strengthen the stretch receptor hypothesis. For example, Wakabayashi extended it by adding diagonal springs for stability and then used the resulting model to investigate several potential locomotion control methods including central pattern generator and stretch receptor control [4]. Using a fairly accurate recreation of the locomotion neural circuit, Cohen et al. have made extensive use of this model to study the use of stretch receptors for creating the phase lag of the wave [28], [5], discount the role that gap junctions play in muscle control [29], and most recently to demonstrate that the worm uses a single gait in liquids of varying viscosity [30].

Several other models have been created that explicitly model the physical body of *C. elegans*. For example, Suzuki et al., used a 12-link rigid segment model to replicate the

Table I
BODY SEGMENT SIZES (ALL NUMBERS IN μm)

i	l_i	r_i	i	l_i	r_i	i	l_i	r_i
1	50	24	**10**	45	40	**19**	35	40
2	10	28	**11**	30	40	**20**	50	40
3	15	34	**12**	55	40	**21**	55	40
4	25	40	**13**	40	40	**22**	30	40
5	20	40	**14**	50	40	**23**	40	38
6	25	40	**15**	40	40	**24**	50	32
7	20	40	**16**	50	40	**25**	100	24
8	40	40	**17**	55	40			
9	30	40	**18**	40	40			

worm's response to touch [31]. Ronkko and Wong, on the other hand, used a three-dimensional particle-based model to explore the worm's swimming and crawling behavior in various substrates including agar, water, and soil [32]. Finally Ferree, Marcotte, and Lockery built a model worm that moves forward at a constant velocity, but changes direction using a neural network that has been trained to replicate chemotaxis behavior [33], [34].

A number of simulators have also been constructed that model aspects of the nervous system without explicitly modeling the body of the worm. The work of Karbowski et al. simulates the neural circuits involved in locomotion [3]. Using a custom simulator, Wicks, Roehrig, and Rankin built a model of the neural network that controls the tap withdrawal response and by systematically analyzing it, were able to derive a possible functional relationship between some neurons in this circuit [35].

III. SIMULATOR

To develop the core of our simulation framework (see Figure III), we chose to use a Java-based high performance 3D game engine called the Java Monkey Engine (JME) [36]. Originally created by Mark Powell and now in its third major revision, JME provides all of the major features found in commercial quality game engines including loading and manipulation of 3D meshes, lighting and shadows, sound effects, animation, and terrain. JME uses a scene graph based API that allows developers to easily modify composed objects in their scene and the game engine to quickly cull branches of the graph during rendering. This makes it both easy to use and exceptionally fast.

At the core of our simulation framework is the Open Dynamics Engine (ODE) [37], which interfaces with JME using JME Physics [38]. ODE is designed to simulate articulated rigid body physics. Objects in the simulation are built from various 3D shapes that are connected to one another by joints. ODE allows users to specify the properties of the objects including weight, surface friction, and center of gravity. Joints can be created between the objects and up to six degrees of freedom are supported.

ODE uses a highly stable, first-order implicit Euler integrator. Although not quite as accurate as a fourth-order Runge-Kutta integrator, it is remarkably fast and with small enough time steps provides very realistic physical approximations. ODE handles contact and friction using a version of the Dantzig Linear Complementarity Problem (LCP) solver that was described in [39]. However, it uses a faster Coloumb friction model to optimize speed. ODE is used in a number of research and commercial robotics simulators including Gazebo [40], Marilou [41], and Webots [42].

A. Physics Model

Because ODE is designed to simulate rigid objects, we modeled the body as set of 25 discrete segments $S_i = \{1, \cdots, 25\}$ (see Figure 2). As a notational convenience we refer to segments by their index number and use the subscript i in equations to denote the segment S_i. Each segment is represented using a 3D box whose width and height are estimated by the radius, r_i, taken from photographs of living worms and length, l_i, are given by the spacing between subsequent muscle cell locations along the worm's anterior-posterior axis (see Table I). These muscle cell locations, which are taken from [43], represent the main points of powered articulation along the body .

The volume of each segment can be calculated by

$$v_i = 4r_i^2 \times l_i \tag{1}$$

and their mass, w_i, is a fraction of the total mass W and can be calculated by

$$w_i = \frac{W \times v_i}{v_{total}} \tag{2}$$

This representation creates non-uniform weights and sizes for the individual sections of the worm. This, in turn, impacts the shape and frictional properties associated with the contact surface between the worm and its environment.

Subsequent segments of the body, S_i and S_{i+1}, are connected to one another by a powered rotational joint, J_k. Like segments, we use the notational convenience of referring to the joint by its index number and use the subscript k to denote joint J_k. These joints have 2 degrees of freedom and have an angle at time t that is represented as a 3D vector $\vec{\theta}_k(t)$. The angular velocity of this joint $\vec{v}_k(t)$, is also represented as a vector with each element being the change in the corresponding angle over time.

The values of $\vec{\theta}_k(t)$ and $\vec{v}_k(t)$ are calculated by the underlying physics engine based on the various forces that are acting on the joint and the segments they connect. This is the topic of the next section.

B. Dynamics

At any given time there are a number of forces that act on the body of the worm. These include gravity, friction, surface tension, inertia, elastic forces from the cuticle, force associated with internal pressure, dampening forces caused

Table II
PARAMETERS USED IN THE SIMULATION

Parameter	Description	Value	Units
L	Total length of the body	1000	μm
W	Total mass of the body	$5 \cdot 10^{-9}$	kg
S	Surface Tension	$2.0 \cdot 10^{4}$	G
F_m^{max}	Max muscle cell force	$6.5 \cdot 10^{-12}$	N
θ_m^{max}	Max muscle cell bend	$\frac{\pi}{10}$	rad
k_s	Torsional spring constant	$20 \cdot 10^{-12}$	$N \cdot m/rad$
k_p	Torsional pressure constant	$0.63 \cdot 10^{-12}$	$N \cdot m/rad$
b	Torsional damping constant	$0.2 \cdot 10^{-12}$	$N \cdot m \cdot s/rad$
μ_l	Coefficient of lateral friction	0.5	
μ_a	Coefficient of axial friction	5×10^{-3}	
λ_f	CPG wavelength	22	joints
Δt_f	CPG update interval	150	ms
λ_r	CPG wavelength	18	joints
Δt_r	CPG update interval	250	ms

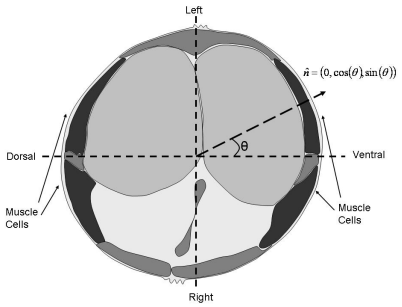


Figure 5. Cross-section showing the location of the muscle cells

by the incompressibility of liquid, and forces exerted by the muscles.

Fortunately, ODE handles gravitational, inertial, and frictional forces that act on the worm by allowing users to specify the mass (for computing gravitational and inertial forces) and the size, shape, and frictional coefficients for each of the segments of the body (for computing frictional forces). Tables I and II list the values that are used in our simulation.

Notable among these are the values for lateral and axial friction along with the value for surface tension. Currently, there are no exact values for the friction between the body of the worm and an agar surface. However, researcher believe that the alae that run the length of the worm's body act similar to an ice skate blade and creates very low axial friction while providing as much as 100 times the amount of lateral friction [2]. In our simulations, we have adopted an axial friction coefficient that is similar to a skate on ice, which is reported to be 0.005 [44] and use a value 100 time this amount for the lateral friction coefficient.

Unlike most organisms, the life of *C. elegans* is dominated by the force of surface tension, not gravity [45]. In fact, estimates for the surface tension experience by the worm are somewhere between 10,000 and 100,000 times the force of gravity [46]. At first, these numbers appear to be outrageously high, however we have recently conducted experiments using a high speed centrifuge, which, to our utter amazement, definitely show that the worms have no problem adhering to agar at forces of over 8,000 Gs. For our simulations, we chose to use a value of 20,000 Gs, which is simulated using a directional acceleration of $20000 \times -9.81 m/s^2$.

The forces that act on the joints are computed by our simulator about 100 times per second during each update of the physics model. The total force applied to a joint J_k at time t is calculated using the following equation:

$$\vec{F}_k^T(t) = \vec{F}_k^M(t) + \vec{F}_k^S(t) + \vec{F}_k^P(t) + \vec{F}_k^D(t) \qquad (3)$$

where $\vec{F}_k^M$ is the force exerted by the muscles, $\vec{F}_k^S$ is the force exerted by the elastic cuticle, $\vec{F}_k^P$ is the force exerted by the interior hydrostatic pressure, and $\vec{F}_k^D$ is a dampening force.

1) Muscle Forces: Muscle forces in *C. elegans* are produced by muscle cells that are attached to the cuticle of the body. These cells are arranged parallel to the anterior-posterior (AP) axis of the body in four rows with two rows on the dorsal and two rows on the ventral side. Figure 5 show a cross section of *C. elegans* showing the location of the muscles cells approximately half way down the AP axis. As can be seen in this figure, the muscle cells are offset by approximately 30 degrees left and right of the dorsal-ventral midline. Muscle cells are, therefore, named based on their AP position in the row and the row's position relative to the dorsal-ventral and left-right mid-lines. For example , the 7th muscle cell in the dorsal, right row is called MDR07. For simplicity, we refer to the set of muscle cells located at joint k as $m_k = \{MDL_k, MDR_k, MVL_k, MVR_k\}$.

For most of the worm's body, this offset has very little effect because the innervation pattern activates both the left and right muscle cells on a single side of the worm at the same time. This is not true in the head where a more complex innervation pattern allows the worm to lift its head

by activating the muscle on the left or right side at the same time.

We accounted for the offset placement of the muscle cells by calculating the force generated by each cell independently and multiplying that force by $\hat{\mathbf{n}}$, which is a unit vector normal to the surface of the worm. The normal vector is easily computed by $\hat{\mathbf{n}} = (\mathbf{0}, \pm\cos(\theta), \pm\sin(\theta))$ with signs being set appropriately according to quadrant location of the muscle cell. The following equation is used to calculate the muscle force applied to joint k at time t:

$$\vec{F}_k^M(t) = \sum_{m \in m_k} a_m \times F_m(\vec{\theta}_k(t)) * \hat{\mathbf{n}} \quad (4)$$

In the equation, a_m is the current activation level, a graded signal, of the muscle based on the inputs from the neurons innervating muscle cell m and $F_m(\vec{\theta}_k(t))$ is the maximum force that the cell can produce given its current length.

To compute $F_m(\vec{\theta}_k(t))$, we used a linear approximation to the Hill equation for the force/length relationship of muscle cells [47].

$$F_m(\vec{\theta}_k(t)) = (0.5k_s - \frac{F_m^{max}}{\theta_m^{max}}) \times |\vec{\theta}_k(t)| + F_m^{max} \quad (5)$$

Here, F_m^{max} is the maximum force that the cell produces at resting length, θ_m^{max} is the maximum angle that the joint can be displaced as a result of the contraction of the muscle and k_s is the spring constant associated with the elastic cuticle. The slope of this line is based on the spring constant of the cuticle such that these forces come to equilibrium when $|\vec{\theta}_k(t)| = \theta_m^{max}$ and both the right and left muscle are fully activated. This, by no means, limits the maximum bend angle of the joint. Both external and inertial forces can cause a joint to exceed θ_m^{max}. The values for these constants (Table II) were chosen based on values from [48] and [4] and were tuned using videos of worms during forward locomotion (see Section IV).

2) Spring, Pressure, and Damping forces: Whereas muscle force causes the body to deviate from it resting state, there are a number of forces that act to restore it. One of these forces is the elasticity of the cuticle, which can be modeled as a simple spring. Below is the equation for the force exerted by this spring:

$$\vec{F}_k^S(t) = -0.8 \times k_s \times \vec{\theta}_k(t) \quad (6)$$

The value for k_s is strongly related to the maximum muscle force and was chosen to be $k_s = 4 * F_m^{max}$. This creates a relationship between these values such that the average force applied over the range of dorsal or ventral muscle contraction is $\frac{1}{2}F_m^{max}$.

Along with the force created by the cuticle, internal pressure of the worm's body also exerts a restorative force. Recent measurements of the relationship between the elastic cuticle and the hydrostatic pressure using a piezoresistive displacement clamp have shown that the restorative force has a cuticle to pressure force ratio of 4 to 1 [49]. We used these finding to normalize the force associated with these two factors. Below is the equation we use for calculating the force associated with internal pressure:

$$\vec{F}_k^P(t) = -0.2 \times k_s \times \theta_m^{max} \quad (7)$$

We explicitly do not model the relationship between the change in volume and pressure as the body bends. We ignored this factor because it has been reported that total body volume does not change significantly over the worm's range of motion [2], [4]. This indicates that pressure acts a constant, not dynamic, restorative property, so we treat it as such.

Lastly, because of the structure and composition of the worm's body, it acts like a fluid-filled shock absorber. So, we apply a damping force to model the resistance that the body has to rapid changes in position using the formula below:

$$\vec{F}_k^D(t) = -b \times \vec{v}_k(t) \quad (8)$$

The value for b was derived through experimentation and is in line with value reported in [4].

Table III
SYNAPTIC WEIGHTS USED IN THE SIMULATION.

Forward				Reverse			
k	ω_k^f	**k**	ω_k^f	**k**	ω_k^r	**k**	ω_k^r
1	0.35	**13**	0.77	**1**	0.10	**13**	1.00
2	0.42	**14**	0.77	**2**	0.20	**14**	1.00
3	0.70	**15**	0.70	**3**	0.30	**15**	1.00
4	0.77	**16**	0.62	**4**	0.40	**16**	1.00
5	0.77	**17**	0.54	**5**	0.50	**17**	1.00
6	0.77	**18**	0.46	**6**	0.60	**18**	1.00
7	0.77	**19**	0.39	**7**	0.70	**19**	1.00
8	0.77	**20**	0.39	**8**	1.00	**20**	1.00
9	0.77	**21**	0.39	**9**	1.00	**15**	1.00
10	0.77	**22**	0.39	**10**	1.00	**22**	1.00
11	0.77	**23**	0.39	**11**	1.00	**23**	1.00
12	0.77	**24**	0.39	**12**	1.00	**24**	1.00

C. Locomotion Control

Because the major focus of this paper is on investigating a realistic physics model of *C. elegans*, we chose to implement a simple CPG for locomotion control similar to the one found in [2]. The CPG works by setting the activation level of the D-type interneuron that drives the B-type motor neurons during forward locomotion, and the A-type motor neurons during reverse locomotion. During forward locomotion this is done by setting the value in the first joint, $k = 1$ at each time t using the following formula:

$$a_{k=1}(t) = \sin(\frac{2\pi}{\lambda_f} \times t) \quad (9)$$

where λ_f is the wavelength for forward locomotion.

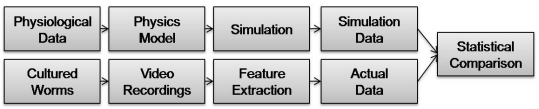


Figure 6. Methodology used to construct and validate the simulator.

At each successive time step $t + \Delta t_f$, the wave is propagated down the body such that for $2 \leq k \leq 24$

$$a_{k+1}(t + \Delta t_f) = a_k(t) \tag{10}$$

The value of Δt_f sets the characteristic frequency of the body wave for forward locomotion (see Table II). Initially, all of the activation levels are set to 0 and between subsequent updates the interneurons are held at their last activation level.

During reverse locomotion, the CPG sets the activation of the interneuron controlling the motor neurons in the last joint and propagates the wave from back to front. The wavelength and frequency of this wave is determined by the parameters Δt_r and λ_r.

The actual activation levels of the muscle cells a_m are dictated by three factors: (1) the activation level of the interneuron at the joint (a_k), (2) whether the worm is moving in forward or reverse, which determines if the A or B-type motor neuron is being activated by the interneuron, and (3) the synaptic strength of the connection between the motor neuron and the muscle cell ω_k^f or ω_k^r. The following equation shows this relationship during forward locomotion:

$$a_m = a_k \times \omega_k^f \tag{11}$$

In a close approximation to the cross inhibition circuit in *C. elegans*, when a_k is positive, the activation of the ventral muscles (MVL and MVR) are positive and the dorsal muscles (MDL and MDR) are 0. However, when a_k is negative, the activation of the dorsal muscles (MDL and MDR) becomes positive and the ventral muscles (MVL and MVR) are 0. The synaptic strength between the motor neurons and muscle cells were determined experimentally using data derived from video analysis of the worm in motion. Table III gives the weights that were found to produce a very close approximation to the gait of the worm during forward and reverse locomotion.

IV. Tuning and Validation

The value of a simulation can only be measured by cross-validating the data it produces against reality. In constructing and validating our simulation, we followed the process outline in Figure 6. The next phase of construction for the simulation was to tune and cross validate its performance against real worms. To do this, we used the *C. elegans* N2 wild type strain, which we obtained from the Caenorhabditis Genetics Center (CGC) at the University of Minnesota.

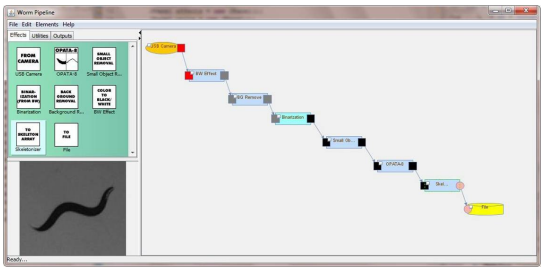


Figure 7. Screenshot of the WormAnalyzer

A. Materials and Methods

Maintenance and culturing of worms was performed as outlined in [50]. Worms were grown on standard NGM Lite plates with OP50 1CGJ *E. Coli* and incubated at $20°C$. The worms analyzed in these experiments were young adult worms (or, a mixture of young adult and adult worms), transferred while in L4 larval stage to freshly seeded OP50 plates the night before filming [51]. Worms were recorded on modified NGM Lite agar plates (3.0g KH2PO4, 0.5g K2HPO4, 2.0g NaCl, and Agar, without the addition of peptone or cholesterol) made specifically for the clarity of the recorded image. Worms were transferred to filming plates via a worm pick and filmed within thirty seconds of transfer, in order to take advantage of the higher locomotion activity exhibited by worms shortly after transfer. We found that this produced the most reliable and longest sustained locomotor activity.

Worms were filmed using a Leica S8 APO Stereomicroscope fitted with a ScopeTek MDC320 digital camera, outputting at a resolution of 1024x768 at 5 frames per second. The microscope was held at the 16x magnification setting during filming, with the illumination mirror angled to obtain images with the highest possible contrast. Individual worms were filmed for five minutes, during which the agar plate was moved manually, to re-position the worm within the microscope's field of view. A total of 487 minutes of raw video in 96 files was collected.

B. Image Processing

After collecting the video, we edited it to remove plate repositioning. Using consumer video editing software, we manually cut the raw footage and produced around 3.75 hours or 50,000 individual frames of forward locomotion and 21,000 frames or 2.33 hours of reverse locomotion.

To analyze this footage, we developed software, called the WormAnalyzer (see Figure 7, that is written in Java and based on the Java Media Framework (JMF). To do this, a set of JMF processors were constructed into a processing pipeline that converted each frame of the raw color video into a set of pixel locations that describes the position of the worms body. The WormAnalyzer software batch processes footage at about 5 times faster than real time. The analysis

pipeline we constructed is similar to that described in [51] with some minor improvements.

The first processor in the pipeline converts the video frames to grayscale and normalizes. This algorithm creates a histogram of color intensity for all pixels within a frame [52]. Using a threshold, we identify which intensities compose the majority of the image. This range of intensities is then converted to an easily identifiable solid color. This greatly diminishes the background noise that, if on the borders of a worm, can resist binarization and small object removal, ultimately skewing the thinning process.

Although the first step in the process converts the image to grayscale and removes most of the background, the second processor in the pipeline converts it to a binary image using a local thresholding algorithm. This algorithm uses a sliding 3 X 3 window to determine whether a given pixel should be converted to black (foreground) or white (background). This processor colors the pixel black if the standard deviation of the intensity of the pixel and its surrounding pixels is greater than the mean of the entire image, or if the mean intensity of the pixel and its surrounding pixels was greater than the background pixel intensity. Because the first processor converts the background to a solid color (white), this phase creates a very accurate binary image.

Next, we remove small objects from the image using a region labeling algorithm that indexes each pixel in the image according to the region that it belongs to. Once all of the black regions are labeled, we remove all of the regions except for the largest. This isolates the worm (colored black) onto a white background. We then used the same method to fill holes in the worms image by inverting the colors such that only the largest white region is maintained (the background).

The resulting binary image, now just the worm and the background, is passed through an implementation of the 8-distance, one-pass asymmetric thinning algorithm ($OPATA_8$) devised by Deng et al [53]. With noise resistance, better connectivity, and less erosion, the $OPATA_8$ implementation more quickly reduces the image to the core skeleton than the previously used triple pass thinning algorithm [54]. The resultant shape can have multiple endpoints but is ultimately trimmed down to a single representative skeleton by selecting the line connecting the two endpoints farthest from one another using the Floyd-Warshall algorithm [55].

The output of this process is a set of text files, which we refer to as body files. Each body file contains one line per frame of video with each line giving a timestamp and the pixel locations of the body of the worm. The number of pixels (or length of the body) is dependent on the size of the worm and also exhibits some variability due to the binarization of the image and subsequent thinning.

Because we wanted to gather statistics based on a nonuniform segmentation of the worms body, it was necessary to identify the location of the worms head. To aid in this process, software was developed to show the first frame of each video file in a directory. To indicate the location of the head, a researcher simply clicked on the head end of the worm. A head tag was then inserted into the corresponding body file that identified the location of the head in the first frame. With the head tag in place, subsequent frames of the video were properly rearranged such that the end point closest to the last head location was identified as the head. This turns out to be a very robust and reliable mechanism if the video being processed was taken at high enough frame rates.

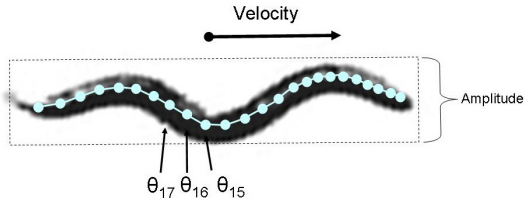


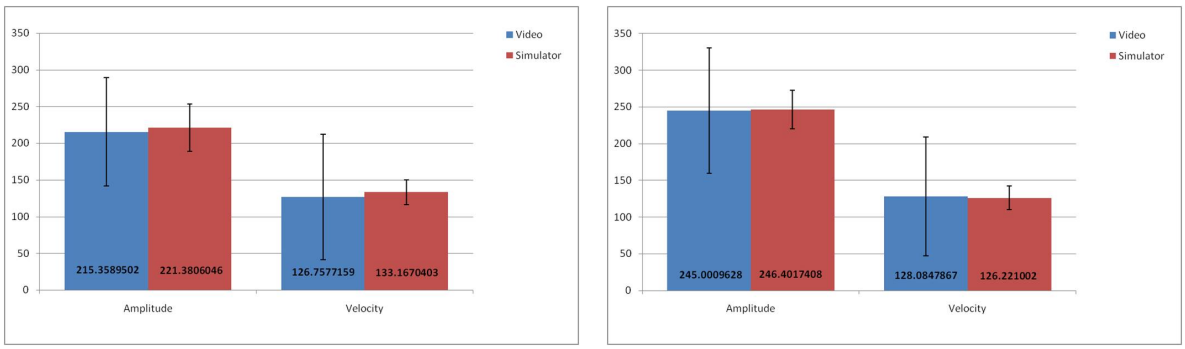
Figure 8. Features of *C. elegans* locomotion

These head-tagged body files are then post-processed in batches. The software takes a directory of head-tagged body files and produces skeleton files that provide a description of the location and position of each segment of the worm. Like the simulator, the size of these segments are not uniform, but are based on the muscle placement as reported in [43]. For convenience, the simulator also creates skeleton files in exactly the same format as the WormAnalyzer. This allows us to directly compare the output from both processes using the same metrics calculated in exactly the same way.

The skeleton files are then processed to extract features of the worms movement using a technique similar to the one reported in [23]. Currently, this software extracts 49 features from the worms motion including the velocity of the centroid of the worm, the amplitude of the worms body, the average angle at each joint location, and the angular to simulated results velocity of each joint (see Figure 8). The software outputs data files that give these features on a frame by frame basis and a set of summary statistics that can be further analyzed using statistical packages.

C. Simulator Data Collection

To collect data from the simulator, the physics model of the worm was instrumented to output the position of each of the joints five times a second to a skeleton file. The CPG was first set to generate forward locomotion and 20 minute long data runs were performed. At the end of each run, the simulator was restarted. We then repeated the experiment while running the CPG in reverse. In total 48,000 data points of both forward and backward locomotion were collected, accounting for approximately 5 hours of runtime. We then processed the skeleton files using the WormAnalyzer.



(a) Forward (b) Reverse

Figure 9. Comparison of locomotion velocity and amplitude of *C. elegans* to simulated results

V. Discussion

Figures 9, 10, and 11 present the results of both the analysis of the video and the simulator. The results for the video are similar with the results obtained in both [3] and [23]. Essentially, it can be seen that the amplitude of the body during forward locomotion is about 21.5% of the length of the body and the instantaneous centroid velocity is about 126 $\mu m/sec$. For reverse locomotion we obtained an average amplitude of 24.5% of the length of the body and a centroid velocity of 128 $\mu m/sec$. These number are similar to the values reported in [23], which gives a forward centroid velocity of 180 ± 30 $\mu m/sec$ and an amplitude of $19.27 \pm 2.34\%$. We believe the discrepancy in velocity is a result of the worms occasionally pausing during our video recordings. We found that obtaining such a slow velocity was very difficult in the simulations.

Figures 10 and 11 show that as the wave propagates down the body that its amplitude decreases as is evident by the bend angles that are produced in these regions. This is similar to the shape of the average flex angles that are reported in [3]. However, we noticed that in our video results, the bend angles increase near the tail of the worm as opposed to continuing to decrease as was reported in [3].

When comparing the video results to the simulated results we see that, with the exception of the angles at the head and tail of the worm, our simulator accurately reproduces the velocity, amplitude, individual angles, and angular velocities of the real worm. We have analyzed the reason for the discrepancy at the head and tail and have concluded that it is caused by data that is not being collected at high enough resolution. The simulation uses a worm of 100 units long with each unit representing $10\mu m$. This makes the second segment of the worm exactly 1 unit in length. When using integers to represent the location of the joints connected to this segment to calculate the angles, it allows only values that are multiples of $45°$. The same error is probably being introduced in the video analysis, but is stabilized by the variability of the worm's length. Overall, these factors effect the analysis, but not the actual behavior of the simulated worm, which by all appearances precisely reproduces the motion of the real organism.

Analytically, the frequency of the CPG used to drive forward locomotion is about 0.3Hz. Even though this is within one standard deviation of the value of 0.36 ± 0.08 Hz for N2 wild type reported in [23], it is still slightly below the expected value. This is entirely expected as the velocity that we produce is lower and the amplitude slightly higher than was reported in that work. We are very confident that changing the values of either λ_f or Δt_f will cause our simulation to reproduce the values they report.

To drive reverse locomotion, the frequency of the CPG is about 0.22Hz. This value is less than the one used for forward locomotion, but a lower frequency was also reported in [3] for reverse locomotion.

We would also like to mention that the process of tuning a CPG to produce the gait of the worm was a very difficult endeavor. There were two issues that we encountered that complicated the task. First, if the weight of the synapse between the interneurons and the motor neurons are uniform, then the model has a tendency to tail whip during forward locomotion. This is caused by a decrease in the wavelength of the wave as it propagates down the body when the worm is in motion, which causes high inertial forces to be exerted. On further investigation we found references to this decrease in wavelength in living worms [22], [30]. To compensate for this effect, we decreased the weight associated with muscles in the rear of the worm, which not only removed the tail whip, but normalized the bend angles to match those observed in the video analysis.

We believe that this indicates that during forward locomotion the worm generates most of its propulsive forces using the muscles located in the anterior. The posterior portions of the worm are likely to apply just enough force to maintain the wave and prevent a loss of energy due to drag. The physiological layout of the worm's musculature seems to support this claim as about 63% of the muscles lie

anterior of the AP midline. We also believe that this has not been observed in other simulations because of the uniform representation of individual segments of the body, the lack of torsional inertial force, and the lack of explicit modeling of the friction caused by variations in surface contact.

The second issue we encountered was the difference in muscle force needed for forward and reverse locomotion. When we first attempted reverse locomotion, we simply reversed the wave pattern. We quickly noticed that the worm was unable to move because the lower concentration of muscles near the tail could not produce enough force. By increasing the maximum muscle force, we were able to get the worm to move in reverse, but then needed to scale back the synaptic weights used during forward locomotion. This could indicate two things. First, it may that the muscles don't need to be flexed with full force during forward locomotion because of their placement. The smaller amplitude and increased frequency used during forward locomotion seem to support this conjecture. It may also indicate that the actual worm does not generate more force while reversing, but is generating it in a different way, driving the movement using the anterior muscles. A more detailed analysis of dynamics of reverse locomotion is needed in order to determine if one or both of these hypotheses is true.

VI. Conclusion

In this paper we presented a new, biologically accurate, three dimensional model of the body of the nematode *Caenorhabditis elegans*. We tuned and validated this model against values derived from both current literature and from analysis of video recordings taken of the worm during forward and reverse locomotion. In the process of performing the tuning, we discovered two new insights into the mechanisms of locomotion employed by *C. elegans*. First, our model shows that worms may derive most of their forward propulsive force from the muscles in the anterior portion of their body with the posterior portions just propagating the wave in an energy minimizing way. Second, we found that in order to perform reverse locomotion that the muscles needed to generate 23% more force than used while moving forward. This may indicate that moving forward actually requires less force from each of the muscles or that reverse locomotion is also predominantly driven using its anterior muscles.

Acknowledgment

The authors would like to thank the CGC for providing the N2 strains used in this paper, Melanie Smith for insightful discussions, Mike Michalek for his assistance in editing the video files, Jason Avery for culturing and videotaping the worms, and the members of the Rand Lab at the Oklahoma Medical Research Foundation (OMRF) for providing their protocols and support on numerous occasions.

The authors gratefully acknowledge support of the Defense Advanced Research Projects Agency under DARPA grants HR0011-07-C-0060. Views and conclusions contained in this document are those of the authors and do not necessarily represent the official opinion or policies, either expressed or implied of the US government or of DARPA.

References

[1] R. Mailler, J. Avery, J. Graves, and N. Willy, "A biologically accurate 3d model of the locomotion of caenorhabditis elegans," in *Proceedings of First International Conference on Computational and Systems Biology and Microbiology (BIOSYSCOM)*, 2010.

[2] E. Neibur and P. Erdös, "Theory of the locomotion of nematodes: Dynamics of undulatory progression on a surface," *Biophysics Journal*, vol. 60, pp. 1132–1146, November 1991.

[3] J. Karbowski, G. Schindelman, C. Cronin, A. Seah, and P. Sternberg, "Systems level circuit model of c. elegans undulatory locomotion: mathematical modeling and molecular genetics," *Journal of Computational Neuroscience*, vol. 24, pp. 253–276, 2008.

[4] M. Wakabayashi, "Computational plausibility of stretch recpetors as the basis for motor control in c. elegans," Master's thesis, University of Queensland, 2006.

[5] J. Bryden and N. Cohen, "Neural control of caenorhabditis elegans forward locomotion: the role of sensory feedback," *Biological Cybernetics*, vol. 98, pp. 339–351, 2008.

[6] W. B. Wood, *The Nematode Caenorhabditis elegans*. Cold Springs Harbor Laboratory Press, 1988, ch. Introduction to C. elegans Biology.

[7] M. Driscoll and J. Kaplan, *C. elegans II*. Cold Spring Harbor Laboratory Press, 1997, ch. Mechanotransduction.

[8] J. Rand and M. Nonet, *C. elegans II*. Cold Spring Harbor Laboratory Press, 1997, ch. Synaptic Transmission.

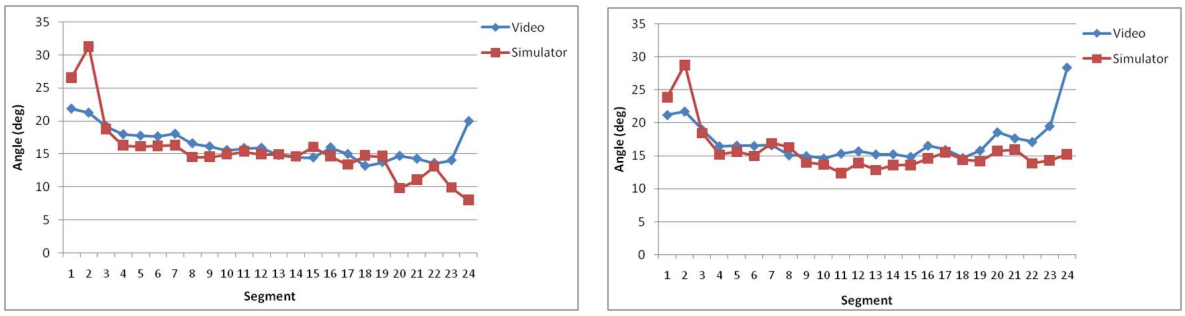
[9] J. White, E. Southgate, J. Thomson, and S. Brenner, "The structure of the nervous system of the nematode caenorhabditis elegans," in *Royal Society of London Philosophical Transactions*, ser. Series B, vol. 314, 1986, pp. 1–340.

[10] J. Richmond and E. Jorgensen, "One gaba and two acetylcholine receptors function at the c. elegans neuromuscular junction," *Nature Neuroscience*, vol. 2, pp. 791–797, 1999.

[11] J. Sulston, M. Dew, and S. Brenner, "Dopaminergic neurons in the nematode caenorhabditis elegans," *Journal of Comparative Neurology*, vol. 163, pp. 215–226, 1975.

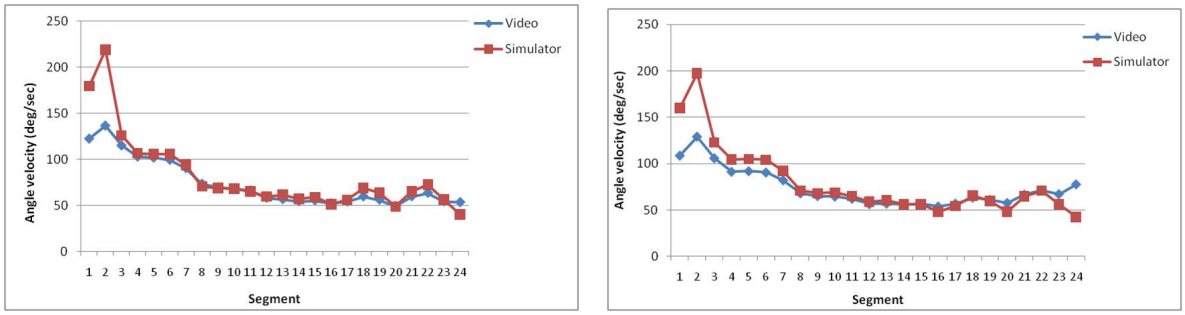
[12] R. Lints and S. Emmons, "Patterning of dopaminergic neurotransmitter indentity among caenorhabditis elegans ray sensory neurons by a tgfβ family signaling pathway and a hox gene," *Development*, vol. 126, pp. 5819–5831, 1999.

[13] S. Srinivasan, L. Sadegh, I. Elle, A. Christensen, N. J. Faergeman, and K. Ashrafi, "Serotonin regulates c. elegans fat and feeding through independent molecular mechanisms," *Cell Metabolism*, vol. 7, pp. 533–544, 2008.



(a) Forward (b) Reverse

Figure 10. Comparison of bend angles in *C. elegans* to simulated results



(a) Forward (b) Reverse

Figure 11. Comparison of bend velocities in *C. elegans* to simulated results

[14] C. Trent, N. Tsung, and H. Horvitz, "Egg-laying defective mutants of the nematode caenorhabditis elegans," *Genetics*, vol. 104, pp. 619–647, 1983.

[15] A. Hart, S. Sims, and J. Kaplan, "Synaptic code for sensory modalities revealed by c. elegans glr-1 glutamate receptor," *Nature*, vol. 378, pp. 82–85, 1995.

[16] M. Goodman, D. Hall, L. Avery, and S. R. Lockery, "Active currents regulate sensitivity and dynamic range in c. elegans neurons," *Neuron*, vol. 20, pp. 763–772, 1998.

[17] C. Bargmann and L. Avery, "Laser killing of cells in caenorhabditis elegans," *Methods in Cell Biology*, vol. 48, pp. 225–249, 1995.

[18] L. Avery and H. Horvitz, "Pharyngeal pumping continues after laser killing of pharyngeal nervous system of c. elegans," *Neuron*, vol. 3, pp. 473–485, 1989.

[19] E. Tsalik and O. Hobert, "Functional mapping of neurons that control locomotion behavior in caenorhabditis elegans," *Journal of Neurobiology*, vol. 56, no. 2, pp. 178–197, 2003.

[20] J. Gray, J. Hill, and C. Bargmann, "A circuit for navigation in caenorhabditis elegans," *Proceedings of the National Academy of Sciences USA*, vol. 102, pp. 3184–3191, 2005.

[21] S. McIntire, E. Jorgensen, and H. Horvitz, "The gabaergic nervous system of caenorhabditis elegans," *Nature*, vol. 364, pp. 334–337, 1993.

[22] J. Korta, D. A. Clark, C. V. Gabel, L. Mahadevan, and A. D. T. Samuel, "Mechanosensation and mechanical load modulate the locomotory gait of swimming c. elegans," *Journal of Experimental Biology*, vol. 210, p. 23832389, 2007.

[23] C. J. Cronin, J. E. Mendel, S. Mukhtar, Y.-M. Kim, R. C. Stirbl, J. Bruck, and P. W. Sternberg, "An automated system for measuring parameters of nematode sinusoidal movement," *BMC Genetics*, vol. 6, no. 5, February 2005.

[24] C. Bargmann and I. Mori, *C. elegans II*. Cold Spring Harbor Laboratory Press, 1997, ch. Chemotaxis and Thermotaxis.

[25] M. Driscoll and J. Kaplan, *C. Elegans II*. Cold Springs Harbor Laboratory Press, 1997, ch. Mechanotransduction.

[26] A. Stretton, R. David, J. Angstadt, J. Donmoyer, and C. Johnson, "Neural control of behavior in ascaris," *Trends in Neuroscience*, vol. 8, pp. 294–300, 1985.

[27] E. Neibur and P. Erdös, "Theory of locomotion of nematodes: Control of the somatic motor neurons by interneurons," *Mathematical Biosciences*, vol. 118, no. 1, pp. 51–82, 1993.

[28] J. Bryden, "A simulation model of the locomotion controllers for the nematode caenorhabditis elegans," Master's thesis, University of Leeds, 2004.

[29] J. H. Boyle and N. Cohen, "Caenorhabditis elegans body wall muscles are simple actuators," *Biosystems*, vol. 94, pp. 170–181, 2008.

[30] S. Berri, J. H. Boyle, M. Tassieri, I. A. Hope, and N. Cohen, "Forward locomotion of the nematode c. elegans is achieved through the modulation of a single gait," *HSFP Journal*, vol. 3, no. 3, pp. 186–193, 2009.

[31] M. Suzuki, T. Tsuji, and H. Ohtake, "A dynamic body model of the nematode c. elegans with touch-response circuit," in *Proceedings of the 2005 IEEE International Conference on Robotics and Biomimetics*, 2005, pp. 538–543.

[32] M. Rönkkö and G. Wong, "Modeling the c. elenags nematode and its environment using a particle system," *Journal of Theoretical Biology*, vol. 253, pp. 316–322, 2008.

[33] T. Ferree, B. Marcotte, and S. Lockery, "Neural network models of chemotaxis in the nematode caenorhabditis elegans," in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds. MIT Press, 1997, pp. 55–61.

[34] T. Ferree and S. Lockery, *Computational Neuroscience: Trends in Research*. Plenum Press, 1998, ch. Chemotaxis Control by Linear Recurrent Networks.

[35] S. Wicks, , C. Roehrig, and C. Rankin, "A dynamic network simulation of the nematode tap withdrawal circuit: Predications concerning synaptic function using behavioral criteria," *The Journal of Neuroscience*, vol. 16, no. 12, pp. 4017–4031, 1996.

[36] M. Powell, "Java monkey engine v2.0." [Online]. Available: http://www.jmonkeyengine.com

[37] R. Smith, "Open dynamics engine." [Online]. Available: http://www.ode.org

[38] O. S. Project, "Java monkey engine physics v2.0." [Online]. Available: https://jmephysics.dev.java.net/

[39] D. Baraff, "Fast contact force computation for nonpenetrating rigid bodies," in *Proceedings of SIGGRAPH 94*, July 1994, pp. 23–34.

[40] O. S. Project, "Gazebo." [Online]. Available: http://playerstage.sourceforge.net/index.php?src=gazebo

[41] AnyKode, "Marilou." [Online]. Available: http://www.anykode.com

[42] Cyberbotics, "Webots v6." [Online]. Available: http://www.cyberbotics.com/

[43] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, "Structural properties of the caenorhabditis elegans neuronal network," 2009. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0907.2373

[44] J. de Koning, G. de Groot, and G. J. van Ingen Schenau, "Ice friction during speed skating," *Journal of Biomechanics*, vol. 25, no. 6, pp. 565–571, 1992.

[45] D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess, *C. elegans II*. Cold Spring Harbor Laboratory Press, 1997, ch. Introduction to C. elegans.

[46] N. Szewczyk, *Basic Biology for Engineers*. NASA Ames, 2003, ch. Unit 4.2: C. elegans Biology.

[47] A. V. Hill, "The heat of shortening and the dynamic constants of muscle," in *Proceedings of the Royal Society of London*, ser. B, vol. 126, 1938, pp. 136–195.

[48] J. H. Boyle, J. Bryden, and N. Cohen, "An integrated neuromechanical model of c. elegans forward locomotion," in *Neural Information Processing: 14th International Conference, ICONIP 2007*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 37–47.

[49] S.-J. Park, M. B. Goodman, and B. L. Pruitt, "Analysis of nematode mechanics by piezoresistive displacement clamp," in *Proceedings of the National Academy of Sciences*, vol. 104, no. 44, 2007, pp. 17 376–17 381.

[50] S. Brenner, "The genetics of caenorhabditis elegans," *Genetics*, vol. 77, pp. 71–94, 1974.

[51] W. Geng, P. Cosman, C. C. Berry, Z. Feng, and W. R. Schafer, "Automatic tracking, feature extraction and classification of c. elegans phenotypes," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 10, pp. 1811–1820, October 2004.

[52] M. Nixon and A. Aguado, *Feature Extraction and Image Processin*, 2nd ed. Academic Press, 2008.

[53] W. Deng, S. S. Iyengar, and N. E. Brener, "A fast parallel thinning algorithm for the binary image skeltonization," *The International Journal of High Performance Computing Applications*, vol. 14, no. 1, pp. 65–81, 2000.

[54] T. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, March 1984.

[55] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (3rd ed.)*. MIT Press., 2009, ch. Chapter 25: All-Pairs Shortest Paths.

A Learner's Technique for understanding Scholarly Articles: An empirical Study

Bee Bee Chua, Danilo Valeros Bernardo
University of Technology, Sydney
Faculty of Engineering and Information Technology
New South Wales, Australia
bbchua@it.uts.edu.au; bernardan@gmail.com

***Abstract*—Scholars write scholarly articles to introduce new concepts and ideas. Unfortunately, not every learner or reader can understand every scholar's work. One reason for this is that the language used in papers is profound, hence many learners find it difficult to cope with the language and understand the ideas put forward in papers. To overcome this problem, our focus in this research is to develop a teaching-based technique to guide learners toward a better way of understanding and learning from scholarly articles. The technique in this paper is validated in case studies with the support of evidence that shows it is a proof of learning concept which has significantly contributed to guiding students to better practice in their learning. In addition, the use of this technique helps to promote educational sustainability by developing students' interest in appreciating and understanding scholarly articles.**

Keywords-scholarly articles; learning technique; education, sustainability.

I. INTRODUCTION

Our recent work [1] demonstrated a practice-based technique to learners. This paper is an extended version which incorporates case studies and compares the results to establish whether the technique can be highly recommended for use.

A scholarly article is defined in many ways. A standard definition describes it as an original research or experimentation written by a researcher or an expert in the field who is often affiliated with a college or university [2]. California State University, Chico [3] conducted a research study to compare and contrast scholarly articles with other article types and established that the language used in scholarly articles is a technical terminology appropriate to the discipline. It is assumed that readers will have a similar scholarly background, but despite this assumption, there is unfortunately no evidence that the process of reading and understanding scholarly articles is easy.

Scholarly articles need to be succinct in order to sustain the reader's interest. Papers are usually circulated within academic institutions and are available to industry because they not only contribute to the body of knowledge, but also to the development of new products and services, new processes and new technology, all of which benefit organizations and society as a whole. They drive innovation and change.

Developing scholarly articles is compelling and challenging. Introducing them in the classroom may frequently be even more challenging, especially if they are to attract students' interest as a support for their learning.

The reasons are essentially twofold. Firstly, some articles are not easily read and understood due to the technical nature of the language used, and secondly, the students' lack of critical research skills disadvantage them in understanding the methodologies used in the development of the articles.

This paper is structured as follows. Section 2 examines the use of state of the art requirement elicitation techniques to understand scholarly papers. Section 3 contains a discussion of related works on learning approaches, and educators' feedback on scholarly articles is covered in Section 4. Section 5 outlines a new technique which is followed by a discussion of data collection methods in Section 6. Section 7 addresses the issue of understanding subject assessment criteria. Section 8 describes the first case study validation of the technique; in section 9, an overview of results from the case studies is presented.. Subsequent case studies are also validated; discussion of their results are in Section 10 and Section 11. Section 12 concludes and updates the direction that research will take in the future.

II. STATE OF THE ART USING REQUIREMENT ELICITATION TECHNIQUES TO UNDERSTAND SCHOLARLY PAPERS

The structure of scholarly papers varies in type and length. Examples of scholarly papers are: 1) research papers, 2) experience reports, 3) short papers, 4) posters, 5) tutorial proposals, 6) tutorials, and 7) panels. A research paper describes original, empirical and theoretical research that is composed of new techniques and tools. Usually, a full-length research paper comprises about eight pages. Some full research papers consist of new interpretations, while others require in-depth case studies for analytical findings.

The three largest groups of people who frequently need to access, retrieve and read scholarly papers are educators, researchers, and students. They read scholarly papers to: 1) conduct new research, 2) collect information, 3) advance knowledge, and 4) collect ideas and translate them into projects.

Although scholarly papers are documents to be read for the importance of their information content, they should also be thought of as undoubtedly significant learning drivers for students, teaching them how to think, reflect and review their knowledge.

To encourage learners to read and understand scholarly articles, some common requirement elicitation techniques from the field of software requirements [4] are widely introduced to group-learners and an individual-learner. Examples of requirement gathering techniques include brainstorming, prototyping, interviews and agile methodology [4] [5] [6] [7] [8] [9]. All have been developed and used to clarify, elicit and confirm requirement needs with users. Such techniques are thought to be worth introducing into the academic environment to promote learning effects by encouraging socialization and interaction between learners. These techniques are useful for group discussion and for promoting group synergy, with its potentially positive effects on student learning.

According to Ambler [5], agile methodology highlights story telling from an unclear scenario. It is effective for eliciting users' or customers' requirements, and is useful for helping users to clarify their knowledge through an implicit method, by putting their ideas into a narrative to help the developer understand their requirements. However, it does not promote a critical review of suggestions or ideas.

As information technology rapidly advances, the availability of learning tools has become increasingly sophisticated. Learning tools [11] [12] provide adequate features and functionality to facilitate better learning. According to Chua et al. [13], these tools support learning activities but cannot replace current strategies or introduce new learning strategies or practices.

Integrating scholarly papers into any learning activity can facilitate the learning process effectively and can stimulate learning by providing interest and excitement. A number of traditional teaching and learning methods fail to explicitly demonstrate how to introduce research into practice-based learning.

Many learning methods focus on theory-based and practical-based components, but very few have integrated research-based components into their learning processes. Few researchers could imagine how the mapping of scholarly papers enables learners to improve their learning performance and even to experience joy in reading them.

III. Related Works On Learning Techniques

The term 'learning' is broad. Buchanan and Huczynski [14] define learning as 'the process of acquiring knowledge through experience which leads to a change in behaviour'. In other words, learning is not just the acquisition of knowledge, but its application by doing something different in the world.

A familiar scenario that has incorporated changes can drive us to learn something new, or adjust to a new way of operating, or to unlearn something. From an organizational learning [15] point of view, learning is associated with two important concepts: the first is the power of knowledge acquisition, and the second is the power of knowledge sharing. Understanding scholarly articles provides readers with knowledge and thus increases their ability to knowledge-share with others [16] [17].

It is therefore important to encourage students to learn through reading scholarly articles. However, integrating these articles in the classroom remains a challenging task for educators.

Acknowledging that this is an issue that impacts learning in the classroom, the focus of this paper is to introduce a learning technique that can assist educators to integrate scholarly articles and case-based learning in their teaching.

Case based learning is not a new concept in education. It is effective, but can be challenging. These challenges have been discussed widely in research that focuses on achieving better learning experiences by recognizing the depth of the subject content while increasing the capacity of the learner to develop skills, including problem solving skills [18] [19] [20] [21].

Case based learning can be conducted either by individuals or by groups. Traditionally, the method involves face-to-face teaching. Although some researchers claim that face-to-face teaching of case based reasoning is one of the most traditional and effective learning methods, it demonstrates a lack of learning innovation. Face-to-face teaching is usually conducted in a classroom environment where one or more learners absorb the concepts or theories directly from an educator. The learner can clarify immediate doubts directly with the educator.

This method promotes a dual learning loop: questions from learners and feedback from educators. The drawback of this approach is that not all learners are able to accept and adapt to an educator's teaching techniques. In particular, the technique used in case based reasoning does not promote student learning through the sharing of ideas and knowledge among individuals in the class. Hence, some students find learning difficult, rather than enjoyable or fun. In the worst case, students can become bored with a single and lengthy case study, and instead of their learning horizons being widened, their thinking narrows to focus solely on the case.

IV. Educators Feedback on Scholarly Articles

We surveyed educators to understand what their aims were for learners who have read scholarly papers. Every educator has different expectations and requirements; for example, some educators provide scholarly articles for learners to read, but their requirement is for learners to summarize the article in their own words. This means writing a short version of the research paper from which the educator can assess the learner's writing and analytical skills.

The process is similar to the 'requirements elicitation' technique.

Some educators, however, do not ask for a summary page but want to know how the learners judge the scholarly articles; in other words, they want to test their evaluative skills. This process is similar to the requirement gathering technique called 'prototyping'.

Some educators want learners to answer questions in response to scholarly articles. The aim is to test their critical analysis in problem solving, and this process is similar to the interview technique. Some educators want learners to discuss what the article is about, a process which is similar to the agile methodology of story-telling. One commonality of these approaches for scholarly papers is managing and eliciting requirements.

V. A New Technique

Our objective in introducing the new technique is to provide learners with a better way of understanding scholarly articles by a combination of processes (eliciting, analyzing, clarifying, reviewing, verifying and validating) so that they can get the most out of papers which can provide them with valuable knowledge.

The inward process of the technique emphasizes how learners' skills (communication, analytical and team skills) can be strengthened, and the outward process of this technique is a knowledge sharing mechanism for others. In addition, this technique can help to strengthen learners' skills by making them: 1) responsible for interaction, 2) accountable for critical review, and 3) empowered to produce innovative ideas and decision making. The diagram in Figure 1 illustrates each process.

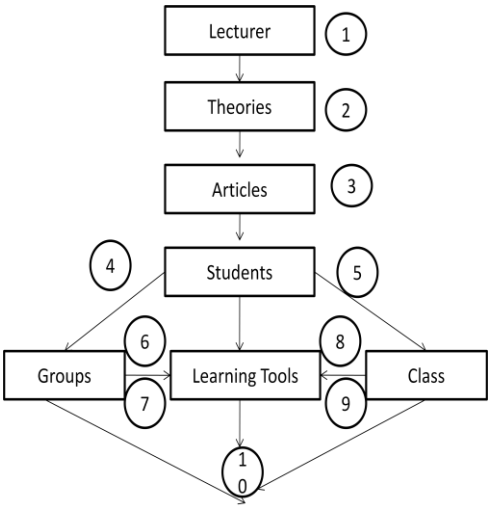


Figure 1. The framework for a practice-based technique on scholarly papers

The steps below show how each individual task is processed.

1. *Delivery by lecturers*

Theories are delivered by lecturers to the class**.**

2. *Imparting knowledge of theories to the class*

Students learn these theories from the lecturer, and their ability to understand concepts is assessed by giving them scholarly articles that relate to the theories discussed in class.

3. *Introducing scholarly articles to students*

Selected articles are distributed to students.

4. *Communicating and sharing of information by students with the aid of the learning tool*

This tool supports and facilitates discussion on the paper's topic.

5. *Class contributing their answers on the learning tool*

All students in the class take part in the discussion using the learning tool.

6. *Group problem solving*

Students in each group perform brainstorming sessions to understand the article and decide the questions to be asked in class for class comments and suggestions.

7. *Uploading questions and answers on the learning tool*

Individual groups upload either open-ended or closed-ended questions in the forum of a learning tool to address problems, concerns and challenging issues discussed in the paper.

8. *Class participation*

The class reads the papers and provides answers based on the questions asked by the group, either quantitatively or qualitatively. Students are also encouraged put related questions to the group.

9. *Individual student participation*

Individual students must answer the questions discussed in the forum.

10. *Group presentation*

Students present their findings in class.

The presentation covers: 1) understanding the paper's content, 2) addressing problems and concerns about the paper, 3) discussing questions and answers posted on the forum, 4) consolidating findings in a summary format, and 5) proposing a strategy, if necessary in relation to the questions asked by other students in the class. Feedback is provided by the class and the lecturer.

Current designs of learning and teaching techniques [11] [12] [13] are useful, especially for widening the range of teaching materials that can be easily understood by students

and that will encourage them to engage in deep learning rather than surface learning. Nonetheless, they lack the ability to elaborate interactively on the students' learning, and no significant evidence was found in the literature review to demonstrate that these techniques provide good support for students in learning scholarly articles effectively.

VI. DATA COLLECTION METHOD

Concern as to how the technique will be validated and the number of case studies needed for such validation is, no doubt, a crucial issue. A framework is proposed to validate the technique in classes and observe step by step how the process works on educators and students. The steps below outline the sequence of the data collection process:

1. Scholarly articles selected by educators.
2. Formation of groups to be decided by educators.
3. Each group is asked to select a scholarly paper.
4. Each group is asked to read and analyse the paper, and then to highlight an important concern that has not been discussed in the paper.
5. Each group is required to upload questions and ask the class for their participation on the learning tool.
6. Each group must summarise class members' feedback on questions and present their discussions and/or answers in class.
7. In a particular week, an anonymous survey will be distributed to students to comment on the technique.
8. Students return the survey to the subject coordinator for data analysis and data interpretation.
9. Three results are revealed. The first result presents an overall statistical rating on closed-ended questions; the second presents comments made by students on questions posted on the forum; and the third result offers qualitative analysis based on the open-ended questions.

VII. SUBJECT ASSESSMENT CRITERA

Students are informed of the marking criteria in this subject. Figure 2 shows the subject assessment criteria for understanding scholarly articles. Students were asked to rank their priorities in understanding scholarly articles. Of the three assessment criteria, seven groups gave the component of research skill the highest mark on innovation and invention (see Figure 3).

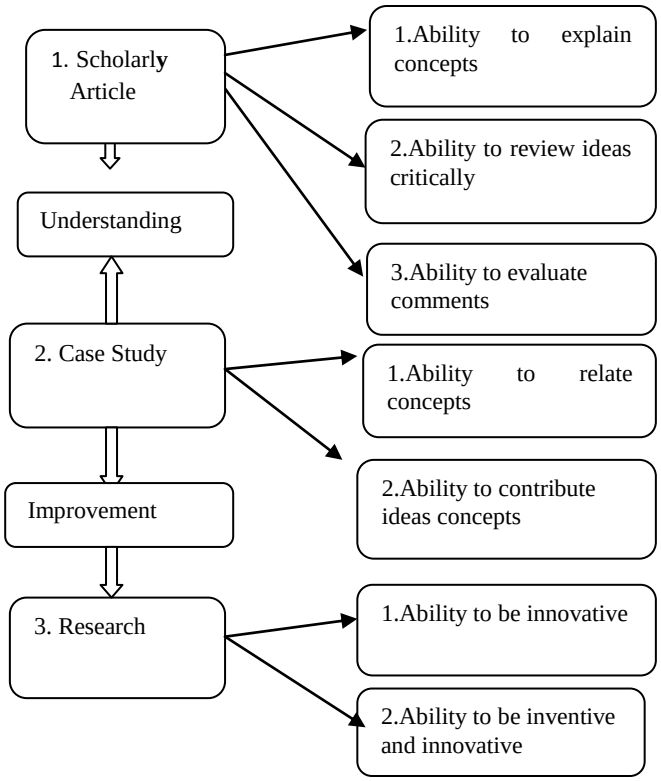


Figure 2. Subject Assessment Criteria

They must have realized that without a good understanding of scholarly papers, it is very unlikely that they would be able to incorporate the research ideas into their discussions on a learning tool

Goal Prioritization by Group	**Class A**					**Class B**				
Group	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Scholarly article Structure	1	2	2	3	2	1	3	2	2	3
Case Study	2	3	3	2	3	3	2	3	3	1
Research Skill	3	1	1	1	1	2	1	1	1	2

Figure 3. Goal Prioritization by Group

VIII. TECHNIQUE VALIDATED IN FIRST CASE STUDY

One author of this paper is a subject coordinator who coordinates a post-graduate subject offered to information technology students. A past survey result showed high and good ratings for the teaching, but not for the subject.

In order to validate our proposed technique, we selected a postgraduate subject having two classes, A and B, as the main focus of the case study and as part of the unit analysis. We carried out an experiment on fifty students from both classes in five weeks and, according to what we observed, data was analyzed from surveys and information that was

posted by students on an online discussion board using an e-learning tool.

Past feedback from many students expressed concern at the difficulty in understanding scholarly articles. Many could not interpret what the authors discussed in the paper. As a result, students did not like the subject or the support materials handed out by the subject coordinator. Rather than re-design the whole subject, the coordinator analyzed all aspects of the learning factors that impacted on the students' learning, and reviewed all processes, including tools and techniques. The learning environment was the first area to be evaluated to discover whether there were any missing or inappropriate resource supports for the students.

The learning tool that was provided to the students provided good functionality and adequate features, according to our observation, and was therefore not believed to be the cause of the problem. As such, the tool was retained. Next, the coordinator reviewed ten different scholarly articles, carefully selected by us, to determine whether they were difficult for students. This review confirmed that there was no replacement of the existing articles, as that was not primary teaching goal. The teaching goal was to encourage students toward deep learning, rather than surface learning and the objective was not, therefore, to change the ten papers being used. Instead, the coordinator revisited the presentation structure, as a result of which it was recognized that it was necessary to re-engineer the presentation process so that the subject matter would be explicitly clear to students, both informatively and intuitively. It was decided to outline any missing steps between the old and new presentation structures, in order to achieve improvement in the subject.

Our objective is to ensure that students are more engaged in their learning and hence we proposed the development of a collaborative interface between students at group and class levels for questions and discussions. This interface acted as a two-way communication process that made groups responsible for posting their designed questions, and the class responsible for feedback on the designed questions.

Fifty students from two classes in one semester took part in the new process. Ten scholarly articles were chosen, on topics ranging from understanding Michael Porter's framework on the five forces to strategic information planning. Papers published by ACM, MISQ and IEEE were the focus. Students listened attentively to the settings for the paper discussion in the first lesson. Each group was made up of five students. Ten groups of five students per group were formed, and each group was given a different paper topic to read, analyze and discuss.

Of the concerns raised, some students were confused about the actual process because it was the first time they had experienced such a technique. A minority of students felt insecure and lacking in confidence because detailed data had to be collected and interpreted in one of the steps, and they had no prior knowledge of research skills.

There were no negative responses from students about the learning process, but acceptance of change was not readily forthcoming when the new technique was introduced.

IX. RESULTS FROM THE FIRST CASE STUDY

The first week of presentations by the two classes went well. Students knew what to do for each paper. They had to: 1) identify a problem issue discussed in the paper (a process equivalent to requirements gathering), 2) contribute their opinions or comments on the paper (a process equivalent to requirements elicitation), 3) ask the class for feedback on questions they asked (a process equivalent to requirements clarification), 4) respond to comments from their classmates (a process equivalent to requirements review), 5) know how to summarize their findings and propose a strategy (a process equivalent to that of requirements changes), and 6) present their data or findings in a class presentation (a process equivalent to requirements traceability).

The presentation structure, the learning tool and the interface for group discussion are the events on which we sought understanding. Students claimed that class A's papers were more difficult than class B's papers. The statistics report showed that class A received more responses than class B, even though the papers were difficult.

We believe that class A students received a high response rate due to the fact that the topic interested them, and thus they focused on that, rather than on the paper's difficulty. The same group of students had to analyze data (feedback) from the class and summarize their findings in one presentation slide. Two of the five questions had to involve a critical review of the research into technology and an analysis of the data collected from their classmates. They were also required to propose ideas for solutions to a particular problem based on their classmates' feedback.

In other words, they had to be able to think of a strategic approach and show why it was useful, thought provoking, innovative and interesting. Most importantly, they were asked to summarize findings from the five questions and to conduct an oral presentation to the class the following week in order to leverage knowledge and knowledge transfer of the topic, ideas and solutions for the class.

TABLE I. Class A and B data with students' responses to the paper

	Class A (16/20)				**Class B (25/30)**					
Paper	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Difficulty	*	*	-	-	*****	-	-	*****	*****	*****

Paper Length	**8**	**15**	**6**	**5**	**24**	**13**	**7**	**8**	**6**	**11**
Responses	**8**	**3**	**2**	**2**	**1**	**4**	**3**	**9**	**4**	**5**

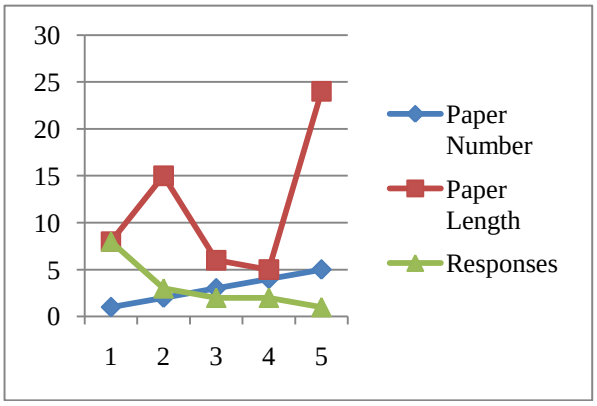


Figure 3. Class A Data with students'responese to the paper.

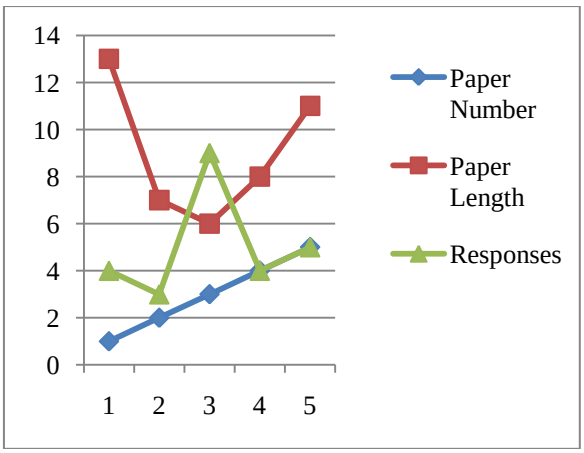


Figure 4. Class B data with students' responses to the paper.

After week five, an anonymous survey was distributed to all students to evaluate their responses to the technique. Forty-one out of fifty students completed and returned the survey. Nine students did not complete it as they did not attend the class. The survey findings are shown in Table 1, Figure 3 and 4.

Difficult papers were rated with an asterisk, indicating that students had difficulty reading them and understanding the scope, and that they had to read them more than once. Before we reviewed the learning process, we were convinced by our students that scholarly papers were too hard to read. We think this is the same belief that drove a similar situation in software development, in which the team always found it difficult to understand some of the users' requirements because they were vague or incomplete. In fact, a well developed process to help developers understand requirements simplifies the situation and makes users' requirements understandable.

This learning technique underpins the process for assisting students to overcome the barrier of reading difficult papers. The aim is to make them realize that academic papers are not complicated or hard to understand. It is a guiding process on the 'how' and 'what' of reading scholarly articles.

We were also keen to know whether students liked the presentation structure. The process for the presentation was to have them read an article, post designed questions and then analyze data from the class feedback and comments from the subject coordinator for an oral presentation. In this question, we were able to gain many valuable insights from students' responses. Most of their comments are similar and we summarized them into four aspects: 1) article topics, 2) paper discussion, 3) questions posted on the forum, and 4) their oral presentation. We were pleased to find that feedback from the students was positive. For the article topics, the words used repeatedly are:

'Topics are current significant, clear and interesting', 'good knowledge', 'It sharpened our thinking', 'Topics are thought-provoking', 'They give us business aspects of a technical field', 'They broadened our knowledge of IT strategies'.

The comments on the paper discussion showed that students felt it was *'informative'*, and that *'team dynamics were unique'*. They agreed that the process involved two-way discussion and they *'enjoyed it'*. They also believed that such discussion helped them *'not only get to know each other better but also able to share their experience and knowledge within the group level and class level'*. On the questions posted on the forum, one student commented that *'questions are a good help to think critically and relate to the paper and real life experiences'*. As for the oral presentation, many students claimed that the purpose was to *'help understand the topic well', 'stimulate discussion in class and feedback from the subject coordinator'*.

Students commented that *'there was a lot of information'* and *'argumentative and critical evaluation'*. They felt that they learned how to *'build oral communication skills, negotiation skills and analytical skills, as well'*.

As a supplementary question, we wanted to know whether students found the presentation structure helpful to their learning; for example, whether it led to better understanding of the scholarly articles. 98% of students agreed that the presentation structure did help them to understand the scholarly papers better. One student offered a comment that was not negative about the presentation

structure, but rather concerned the length of the paper. He felt that some articles were slightly longer than others and thus took longer to read.

Another student believed that some students' answers in the forum discussion showed a lack of clarity – either their answers were incomplete or the meaning was not clear – and it would have been better if they had provided resource links to justify their findings clearly from journals or books.

X. TECHNIQUE VALIDATED IN SECOND CASE STUDY AND ITS OUTCOME

One semester later, the same technique was validated in the same subject. The total number of students enrolled was fifty and each group had ten students. They were given scholarly articles to read and told to use the framework in Figure 1 to assist their understanding At the end of the teaching semester, students were asked to complete a survey designed by the subject coordinator. Students' responses in the result findings (see Table 2 and Figure 5) are similar to those of the first case study.

TABLE II . Class C data with students' responses to the paper.

	Class				
Paper	**1**	**2**	**3**	**4**	**5**
Difficulty	*	*	*	*	*
Paper Length	**7**	**8**	**15**	**10**	**11**
Responses	**9**	**10**	**8**	**10**	**10**

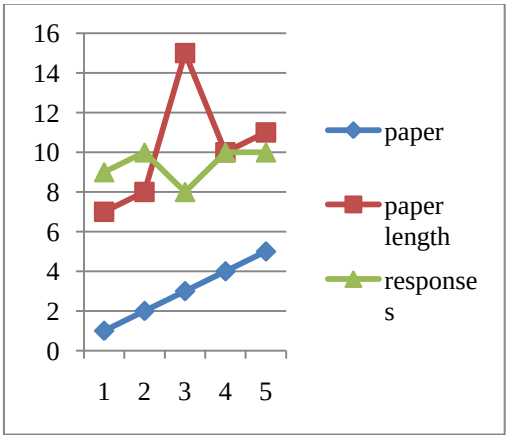


Figure 5. Class C data with students responses to the paper.

Some constructive comments were made in this semester, particularly in relation to the questions posted on the forum, and their oral presentation. Two students commented that the questions posted on the forum by groups analyzed them quantitatively, which did not provide useful insights to the paper topic. Ideally, it would be helpful for groups to provide in depth answers.

XI. TECHNIQUE VALIDATED IN THIRD CASE STUDY AND ITS OUTCOME

It is highly recommended that the technique should be cross validated in different subjects in order to evaluate its results. In another faculty, a research-based subject with heavy emphasis on scholarly articles did not receive a good subject rating, hence the subject coordinator wanted to seek subject improvement. He agreed to use the technique for a trial period during one semester to see whether this would help to improve his subject rating level. He was interested to discover whether the length of scholarly articles affected students' ability to read and understand.

In total, 20 students were enrolled in the subject (Class D). Although the enrolment was not large, the number of students seemed sufficient for us to analyze the results, as long as they were new students learning how to read and understand scholarly articles for the first time.

Fifteen students took part and completed surveys. The information in the returned surveys enabled us to explicitly investigate whether there was any validity threat to the technique. Not to our surprise, the students' feedback was similar to that of students in the first subject. The following Table 3 and Figure 6 illustrate the Class D data.

TABLE III . Class D data with students' responses to the paper.

	Class C (15/20)				
Paper	**1**	**2**	**3**	**4**	**5**
Difficulty	*	-	*	-	*
Paper Length	**9**	**15**	**7**	**6**	**14**
Responses	**3**	**2**	**5**	**2**	**3**

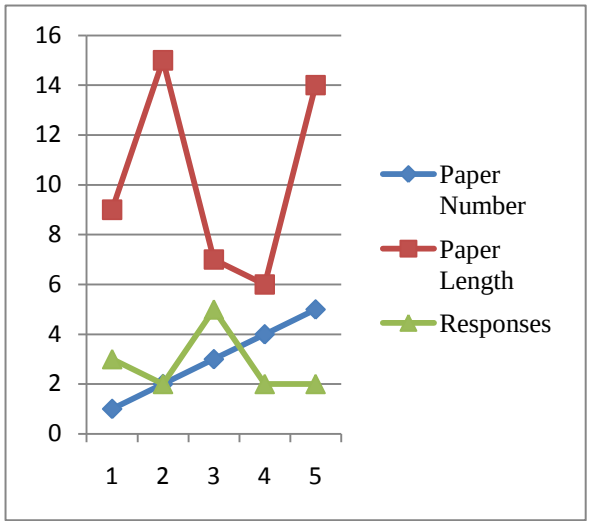


Figure 6. Class D data with students responses to the paper

The results shown in this table and diagram clearly identified to the subject coordinator that there is no significant evidence that students' difficulty in understanding scholarly articles is due to the length of the paper. Three students mention that papers 1, 3 and 5 are difficult despite their length and size. Paper 1 has 9 pages, paper 3 has 7 pages and paper 5 has 14 pages. The most highly rated by students is paper 3. Five students feel that it is difficult.

In the survey, we asked students to comment on the usefulness of the presentation structure. Fifteen students agreed that the process of presentation really helped them to better learn the concepts and theories discussed in the papers. One student commented that the presentation can be time-consuming but is nevertheless thought-provoking.

In order to establish the technique's reliability and effectiveness, it must be validated in more than one case study. The more case studies involved in the validation, the more accurate and reliable the technique can be considered to be.

XII. CONCLUSION AND FUTURE WORKS

Existing case studies reveal that the technique can be applied effectively in research-based and coursework-based subjects in which students might be experiencing difficulty in understanding scholarly articles. The technique appears to be convincing enough to be suitable for use in small classes.

Strategies discussed in this paper are twofold. The first strategy was to observe the technique introduced into the subject and to see the effect on students and their ability to accept difficult scholarly papers. The second strategy was to conduct a survey to measure student satisfaction with the technique.

From the survey, we see that the technique is successful, in particular from the positive comments showing that many students like the technique, and from the relationship between the questions and answers.

Two research strategies were proposed in case studies [22] in order to review the technique to ensure that it is practical and sustainable. We were not simply looking for techniques to assist students to overcome their learning problems; we were also concerned that our technique could be easily used and adapted by educators anytime, anywhere and for any subject.

Our future research study will seek to validate this technique in large classes and in programming subjects, to establish whether it is suitable to use in such contexts. Many concerns remain to be addressed: for example, is this technique able to support a large class of, say, 600 students? Is a learning tool a necessary aid for supporting resources and setting up a forum discussion? What are the limitations of this technique? These questions will roll into the next phase of our research investigation, which will be more in-depth and analytical.

REFERENCES

[1] B. B. Chua and D.V. Bernardo, "Introducing Scholarly Articles: A way for Attaining Educational Sustainability:, In Proceedings of the 2nd International Conference on Mobile, Hybrid, and On-line Learning , pp. 111-115, 2010

[2] Anonymous. What is a scholarly article or book? Available: http://instructional1.calstaela.edu/tclim/definition-boxes/scholarly_article.htm Accessed on 25th November 2010.

[3] California State University, Chico Meriam Library. Available: http://www.csuchico.edu/lins/handouts/scholarly.pdf Accessed on 25th November 2010

[4] I. Sommerville, "Software Engineering". Addison Wesley, Wokingham, UK. 1983.

[5] S. W. Ambler, "Agile Modelling: Extreme Practices for eXtreme Programming and the Unified Process". John Wiley and Sons, New York, NY. 2002.

[6] A. Cockburn and J. Highsmith, "Agile software development: The people factor". IEEE Computer, Vol. 34, No 11, pp.131-133. 2001.

[7] G. Kotonya and I. Sommerville, "Requirements Engineering Processes and Techniques". John Wiley and Sons, New York, NY. 1998.

[8] G. Kotonya, and I. Sommerville, "Requirements engineering with viewpoints". Software Engineering, Vol. 1 No.11. pp.5-18. 1996.

[9] R. Vonk, "Prototyping: The Effective Use of CASE Technology". Prentice Hall, New York, NY. 1990.

[10] R. R. Young, "Effective Requirements Practices". Addison-Wesley, Boston MA. 2001.

[11] L. S. Kheong, "Framework For Structuring Learning In Problem-Based Learning", http://pbl.tp.edu.sg/Understanding%20PBL/Articles/lyejayartna.pdf. Accessed on 30th December 2010

[12] S. Clarke, R. Thomas, and M. Adams, "Developing Case Studies to Enhance Student Learning", , http://crpit.com/confpapers/CRPITV42Clarke.pdf Accessed on 20th January 2011

[13] B. B. Chua and L. E. Dyson, "Applying the ISO 9126 model to the Evaluation of an e-Learning System". In Proceedings of the 21st ASCILITE Conference, pp. 184-190. 2004.

[14] D. Buchanan and A.Huczynski, "OrganizationalBehaviour". Prentice Hall, London. 1995

[15] R. Hussein and J. Goodman,"Leading with Knowledge: The Nature of Competition in the 21st Century". Sage Publications, Thousand Oaks, CA.1998

[16] H. Thomas, T. Davenport, and L. Prusak, "Working Knowledge: How Organizations Manage What They Know". Harvard Business School Press, Boston MA. 1998.

[17] P. Drunker, "The Coming of the New Organization". Harvard Business Review, pp.66-77. 1988.

[18] D. W. Aha, "Case-Based Learning Algorithms". In Proceedings of DARPA Workshop on Case-Based Reasoning, Morgan Kaufmann, San Mateo, CA, pp. 147-157. 1991.

[19] C. Cardie, "Using Decision Trees to Improve Case-Based Learning". In Proceedings of the Tenth International Conference on Machine Learning, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.30.1899&rep=rep1&type=pdf., Accessed on 6th January 2011

[20] C. Drummond, "Using a Case Base of Surfaces to Speed-Up Reinforcement Learning". In Proceedings of the Second International Conference on Case-Based Reasoning Research and Development, Springer-Verlag, London, pp. 435-444. 1997.

[21] J. L. Kolodner, "Case-Based Learning". Morgan-Kaufmann, San Mateo, CA. 1993.

[22] R. K Yin. "Case Study Research: Design and Methods", 4th ed, Sage Publications, Thousand Oaks, 1994.

Visual Instrument Guidance in Minimally Invasive Robot Surgery

C. Staub, G. Panin and A. Knoll
Robotics and Embedded Systems
Technical University Munich
Munich, Germany
{*staub* | *panin* | *knoll*}*@in.tum.de*

Robert Bauernschmitt
Department of Cardiovascular Surgery
German Heart Center Munich
Munich, Germany
bauernschmitt@dhm.mhn.de

***Abstract*—Surgical tool tracking is an important key functionality for many high-level tasks such as the visual guidance of surgical instruments or automated camera control. Readings from robot encoders and the kinematic chain are usually error prone in this kind of complex setup, but still allow for a coarse pose estimation of the instruments in image space. This information can be utilized to (re-)initialize image-based tracking in case of tracking loss and supervise the tracking process. Accounting for the difficult environmental conditions in surgery, the choice of an appropriate tracking modality is important. We have chosen the Contracting Curve Density algorithm (CCD) that maximizes the separation of local color statistics along the contour of a model in contrast to the background. As an application example, the visual guidance of laparoscopic instruments under trocar kinematic is presented.**

Keywords-robotic surgery; minimally invasive surgery; instrument tracking; visual guidance.

I. INTRODUCTION

This paper is an extended version of a conference paper referenced in [1]. It details the formerly announced tracking of surgical instruments and proposes an approach that combines both pose prediction using kinematically derived data and image-based tracking. The extracted position of the instrument is then utilized for visual guidance, a prerequisite for many automation scenarios.

Endoscopic surgery is a challenging technique and has had significant impact on both patients and surgeons. Minimally invasive surgery (MIS) techniques avoid large cuts and patients profit from less pain and collateral trauma. Therefore, the time of hospitalization and the infection rate can be reduced. Unfortunately, surgeons have to cope with increasingly complex working conditions. Long instruments which are unfamiliar and sometimes awkward to operate for the surgeon, are used through small incisions or ports in the body of the patient to perform the intervention. In contrast to the conventional open surgery, visual access to the internal operating scenery is not feasible. This constrains the visual perception to an endoscopic view without an intuitive depth perception or hand-eye coordination. The introduction of telemanipulators, such as the daVinci^{TM}machine [2], has overcome these limitations and is a remarkable example of the ongoing research. The instruments can now be controlled remotely by a surgeon sitting at a master console, which can be placed somewhere in the operation theater. A stereoscopic endoscope provides a 3D view on the situs and improves the perceptual limitations of flattened images. The master console is equipped with sophisticated input devices and provides an intuitive handling of the surgical instruments (Cartesian control without any chopstick effect). The robots at the slave system offer as much freedom of movement as the surgeon's own hand would do in conventional open surgery. Also immersiveness is often improved by means of haptic feedback.

Recently, automation of error-prone and recurrent (sub-) tasks that yield to the quick fatigue of surgeons and noticeable account for a higher overall surgery time have drawn the attention of researchers. Given that knot-tying occurs frequently during surgery, automating this challenging subtask is tackled by several groups (e.g., [3]–[5]). Furthermore, techniques for assisting the surgeon with visually guided instruments (see [6]–[9]) and autonomously navigated endoscopic cameras have been developed (e.g., [10], [11]). Visual servoed instruments are a promising approach in robot-assisted surgery to introduce autonomy and to overcome intrinsic system limitations, often caused by calibration problems. Since visual servoing uses feedback from one ore more cameras to guide a robotic appendage, robust tracking of surgical tools is of particular interest for this kind of application. Also for the reason of documenting and benchmarking surgical interventions, and to anticipate potential mistakes in the surgical workflow, modeling and analyzing surgical procedures has become an active field of research, whereat instrument tracking plays an important role (e.g., [12], [13]).

Despite the manifold of challenges in minimally invasive surgery and the above mentioned achievements in partially autonomous navigation and manipulation, the visual identification, segmentation, and tracking of operated surgical tools during surgery is a crucial requirement for developing techniques that assist the surgeon. As most of the methods require position information of the surgical instrument, a robust and precise automatic detection is

the first step towards higher level functionality. In this paper, we present an approach that allows for a markerless tracking of surgical instruments and its application to visual instrument guidance.

In literature, many of the proposed instrument tracking approaches rely on image processing techniques that use either pure color information or additional geometrical knowledge. Wei et al. [10] analyzed the typical color distribution in laparoscopic images to identify an adequate color for optical markers that are attached to the distal end of the instrument. The marker is segmented in HSV color space and background noise is filtered at a rate of 17Hz. Uckert et al. [14] includes additional shape information about the shaft to fit a bounding box to the color-classified pixels. In order to cope with the typical camera distortion of endoscopes, two different shapes are used: a trapezoidal for near-field cases and a rectangular for far-field cases. In [15], it was taken advantage of the metallic appearance of the shaft to track gray regions by joint hue saturation color features. A seeded region growing method was implemented, operating at 13fps. The fulcrum is estimated with a series of images in order to project an approximated instrument direction and shape into the image. Voros et al. [16] also reduce the search space by considering the insertion point of the instrument. At the beginning of the procedure, the fulcrum has to be visible in the image and is marked with a "vocal mouse". They state that any kind of surgical instrument can be detected since no color information is used, but only the gradients of the instrument edges, constrained by the incision point. To enhance the computation speed, the image resolution is reduced to 200×100 pixels. The precision of the predicted tip position ranges around 11 pixels. The *Center for Computer Integrated Surgical Systems and Technology* (CISST, Johns Hopkins University, Baltimore) tracks the articulated DaVinci^{TM}instruments. Burschka et al. [17] used template images of the instrument to detect the position of the forceps in stereo images, enriched with additional information and orientation information derived from the trajectory provided by the robot. The method works in real-time, but they report that the kinematic data suffers from significant rotational and translational errors. More recently, the CISST reported a general purpose articulated object tracker [18] and demonstrated its application to surgical scenarios. The geometry and kinematics of the objects have to be known a priori. The appearance of different body parts is modeled by a class-conditional probability and compared with the image after rendering the target object geometry. So far, images are hand segmented to train the appearance model and computation time is around $5sec$ per frame at a resolution of 640×480.

The remainder of this paper is as follows. In Section II, the underlying hard- and software that is used for all experiments is introduced. Section III outlines our tracking approach that combines servo readings from the robot as well as image analysis to robustly track surgical instruments in image space. In Section IV, the tracking output is applied to the visual guidance of the instruments. Finally, in Section V, experimental results are presented. Conclusions are drawn and future work is outlined in Section VI.

Figure 1. **Hardware Setup.** Ceiling mounted robots with surgical instruments and master console

II. Robotic System

Several works that are engaged with computer vision aspects in robot-assisted surgery are drawn on a simplified environment. Either the system lacks an endoscopic camera (that usually suffers from strong distortions) or the evaluation was performed within an unrealistic environment. In the majority of cases dimensions of the workspace or distances between camera and instrument are incorrect due to a missing multi-arm setup or port-kinematics. Therefore, the findings of this research project have been assessed within a realistic scenario of robotic surgery [4].

A. Research platform for MIS

As illustrated in Figure 1, the slave manipulator of the system consists of four ceiling-mounted robots which are attached to an aluminum gantry. The robots have six degrees of freedom (DoF) and are equipped with either a 3D endoscopic stereo camera or with minimally invasive surgical instruments, which are originally deployed by the DaVinci^{TM}system. The surgical instruments have 3DoF. A micro-gripper at the distal end of the shaft can be rotated and adaption to pitch and jaw angles is possible. Through the aid of a magnetic clutch the instruments can be interchanged quickly for better handling. The mechanism will also disengage the instruments if forces beyond a certain level are exerted and prevents damage in case of a severe collision. Forces exerting on the instruments are measured by

strain gauge sensors and fed back to the operator by means of haptic devices. The master-side manipulator is equipped with a 3D display, some foot switches for user interaction (such as starting and stopping the system or executing the piercing process) and with the main in-/output devices, two PHANToM^{TM}haptic displays. The devices are used for 6DoF control of the slave manipulator, but also provide 3DoF force feedback derived from the measurements at the instruments. The control software of the system realizes trocar kinematics, whereby all instruments will move about a fixed fulcrum after insertion into the body.

B. Distributed Software Environment

The multi-tier software architecture of our system is distributed over 3 standard PC's: a *simulation and control* PC, a *vision* PC (equipped with a NVIDIA^{TM}Quadro FX 580 graphics card) and one computer is connected to a *CAN* network (cp. Fig. 2). The commands for the servo-motors that control the joints of the instrument as well as the data that is provided by the amplifiers of the strain gauge sensors are communicated between the simulation PC and the PC that is connected to the CAN network. The GUI of the simulation environment comprises an interface to a 3D model of the scene, which can be manipulated in real time. Parameters of each model can be adjusted and joint angles of the robots can be altered this way. New trajectories can be generated by means of a key framing module, incorporating a collision detection. On one hand, joint data is directly sent to the robot hardware, on the other hand the poses of the instruments and the robots are synchronized with the "Vision PC" for further application in image analysis. For this reason, enough computing power can be provided for image analysis, i.e., instrument tracking, visual servoing or augmented reality. Most of the image processing tasks run in individual threads that have access to an image database, which holds up-to-date images provided by the stereoscopic endoscope.

C. System Calibration

Unfortunately, many possible error sources contribute to a comparably high aberration between the real-world hardware and the underlying CAD models of the simulation environment. Camera calibration, exact mounting of the surgical instruments (concerning the magnet coupling) and even the instruments itself introduce quiet large errors. For instance, the flexibility and play of the carbon fiber shaft of the instruments and the gripper at the distal end may vary approximately $\pm 1.5cm$ (cp. Fig. 3(c)). Furthermore, the ceiling mounting of the robots afflicts several intrinsic aberrations, such as variations in the dimensions of the elements and errors of mounting angles. Since all errors sum up, the exact Cartesian position of the distal end of the instrument deviates from the emulation. In order to minimize all intrinsic errors and to establish the

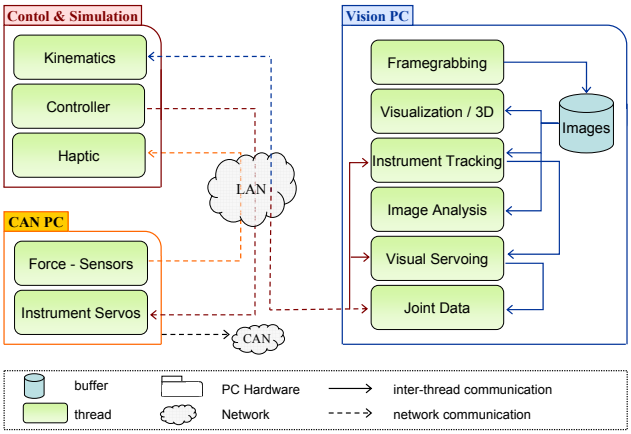


Figure 2. **Software Architecture.** The software of the system is distributed across 3 PC's that communicate via network connections

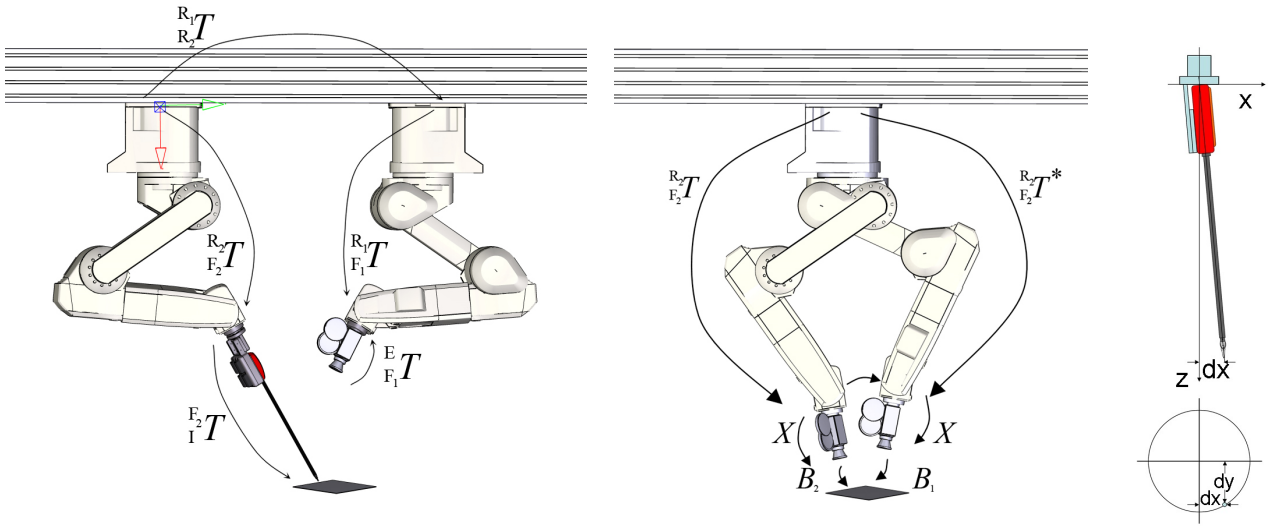
transformations between the individual system components (such as the instrument, different robot bases, etc.) a precise calibration has to be performed. However, an important issue for the acceptance of robotic systems in the operating theater are pre-calibrated components to avoid complicated or long-lasting procedures during an intervention or ahead.

As mentioned above, the robots are mounted on a gantry, assembled of profiled girders. Particularly the coplanarity of the robot's base relative to its attachment cannot be guaranteed and is hardly to be measured. In order to overcome intrinsic variations of the single aluminum elements and errors of mounting angles, a calibration between each of the robot basements is performed.
To align the basements of two robots R_1 and R_2 we employ the following error model:

$$ {}^{0}_{R_1}T \cdot {}^{R_1}_{C}T = {}^{0}_{R_2}T \cdot {}^{R_2}_{C}T \quad (1) $$

In this equation ${}^{0}_{R_1}T$ is the position of the base of the robot R_1, expressed in global coordinates. In order to measure the relative displacement between the robots a calibration frame C in global coordinates is defined and the position and orientation of this frame is measured in local coordinates of each robot. The frame can be replicated by mounting a precisely manufactured calibration trihedron of known size on the flange of both robots. A number of points $M = (p_1, \ldots, p_i)$ are labeled on a checkerboard calibration plate that is positioned in-between the robots, reachable for all four arms (figure 3(a)). The trihedrons of the robots R_1 and R_2 are then driven to all points and the corresponding relative transform (e.g., ${}^{R_2}_{C}T$) can be determined.

Mounting displacements of the robots are not the only source of errors in the system. If we go further down, we find that also the attachment of the instruments bears



(a) System Calibration Overview (b) Hand-Eye Calibration (c) Excentricity

Figure 3. **System Calibration.** Figure 3(a) depicts a schematic overview of the kinematic chain. The hand-eye calibration is exemplified in illustration 3(b). Figure 3(c) shows excentricity and play of the instrument shaft.

certain variances. The magnetic clutch as well as the mechanical fit and the flexibility of the carbon fiber shafts results in a quite high aberration. In case of the endoscopic camera a hand-eye calibration solves this problem [19]. One way to calculate the displacement of an attached endoscope with respect to the flange of the robot, is to solve ${}^{R_2}_{F_2}T \cdot X = X \cdot {}^{R_2}_{F_2}T^*$ (compare Fig. 3(b)). Regarding the surgical tools, this method would introduce two issues in the context of medical procedures: on one hand it is difficult to create a calibration pattern which can be precisely reached by the forceps or attached to the shaft. On the other hand, it is challenging to perform the calibration in the sterile environment of an operating room. The proposed method allows a pre-calibration of every instrument, which can be applied to the system previous to the intervention. To compensate the excentricity, an approximation which simplifies the calculation and applies only to small angles is used. An aberration dx and dy from the center will lead to a positional error or approximately $\sqrt{dx^2 + dy^2}$. The parameters shown in Fig. 3(c) can be found by positioning the instrument over a planar surface with the z-axis of the robot's tool system normal to the surface. By rotating the end effector about $360°$ a circular path is described and the relevant parameters can be determined. In order to compensate for this excentricity, the found correctional transformation has to be applied to the end effector prior to the calculation of the inverse kinematics of the robot.

State of the art endoscopes offer physicians a wide-angled field of view which is imperative for minimally invasive interventions. In order to determine the projective parameters of the camera system a calibration procedure is to be performed a priori.

III. Instrument Tracking

The tracking of surgical tools is particularly challenging due to the changing appearance of the background (e.g., background movement through organs, non-uniform and time-varying lightning conditions, smoke caused by electro-dissection and specularities), but also due to the partial occlusion of the instrument and body fluids that may change the appearance of the instrument itself. In many cases of surgical tool tracking the tracking is constricted to a sequential "frame-by-frame detection" (also referred to as *detection*), rather than including a motion model. Accordingly, no optimization of the configuration space or pose prediction is performed over time.

A. Instrument Tracking Supported by Kinematic Prediction

In a Bayesian *prediction-correction* context, the state of the object is updated by integrating posterior statistics and therewith knowledge about time-depending characteristics of the movement. This "intelligence" within our tracking pipeline is provided by a Kalman filter [20] that is running on the output of a contour tracker, known as contracting curve density algorithm (CCD), based on the separation of local color statistics (see [21], [22]). The separation is performed between the object and the background regions, across the projected shape contour of a CAD model under a predicted pose hypotheses. An overview of the process flow is given in Figure 5.
Tracking always involves a detection step to initialize the system in the very first frame or after encountering a

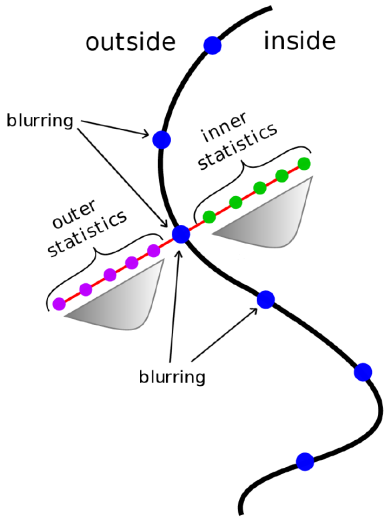


Figure 4. The CCD algorithm tries to maximize the separation of color statistics between two image regions. The algorithm first samples pixels along the normals for collecting local color statistics.

track loss. Instead of simply relying upon visual data, we take an estimated object pose, derived from the kinematic measurement of robot sensor readings. The precision of this approximation is limited due to the absolute accuracy of our system (and the performed calibration).

The idea of integrating joint angle measurements for tracking purposes was e.g., also applied by Ruf et al. [23] to track a polyhedral tool and simultaneously adapt inaccuracies in the static calibration of the robot. To restrict the initial search from the first frame to a specific region is computational more efficient than a complete image analysis and can also be considered from a biological point of view: Biologically inspired algorithms seek to direct the attention rapidly towards a region of interest, using an attention-based type of filter, and only process a smaller amount of the visual input data [24]. *Bottom-up* approaches compute visual salient features, such as regions of high contrast, local scene complexity or high scene dynamics. The second type of visual attention is often referred to as *top-down* attention, as the attention is controlled from higher areas of cognition. Kinematic measurements, which are fed to the visual information processing by another software component (thus, a higher area of cognition), can guide the attention directly to a region of interest

B. Model Building

Our system is equipped with the EndoWrist™needle driver tools that are originally deployed with the DaVinci™system. The instruments are composed of a long shaft, a wrist joint and two brackets. It is represented as a polygonal mesh model (cp. Fig. 6) with 6DoF (3 rotations, 3 translations) by a 4×4 transformation matrix in our simulation environment. In order to represent the instrument in 2D image space, we build a rectangular model with rounded edges at the distal end and neglect the brackets. As already mentioned, CCD maximizes local color statistics (object vs. background) along the model contour. More detail is given in Section III-C. Only three of the object edges can be used for normal contour point sampling and collecting statistics for the CCD algorithm. The fourth edge has to be neglected, since it is not an exterior edge of the shaft and therewith no color separation between model and background is possible. The inclusion of the edge would yield to irrepressible shifting of the model alongside of the shaft.

The instrument's 6 pose parameters are reduced to a planar roto-translational pose s with scale h and rotation θ by projecting the 3D model into the image plane.

$$s = (t_x, t_y, h, \theta) \tag{2}$$

An important aspect of the proposed approach is the use of pose estimates, derived from the kinematic chain, that are fed to the tracking pipeline prior to the visual tracking. Instead of referring the tracking parameters s to global image coordinates, we align the tracking frame with the estimated orientation and position of the instrument T_{ref} in every cycle (cp. Fig. 6). Therewith, the tracking is performed in a *local* coordinate system that is axis aligned with the frame of the instrument model. Since the uncertainty of the state

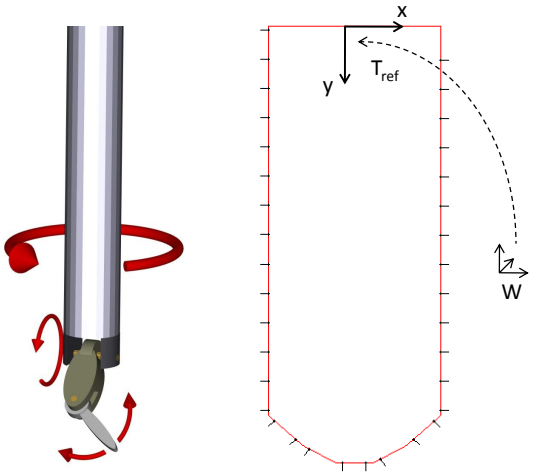


Figure 6. Instrument CAD model and tracking contour model of the shaft with sampling normals.

hypothesis is represented by a squared covariance matrix (with the dimension of the state), we can now alter the matrix and set a higher confidence in the direction of the shaft (the y-axis). This anticipates an uncontrolled sliding of the model alongside of the instrument shaft. The entries in the covariance matrix are found empirically for all DoFs.

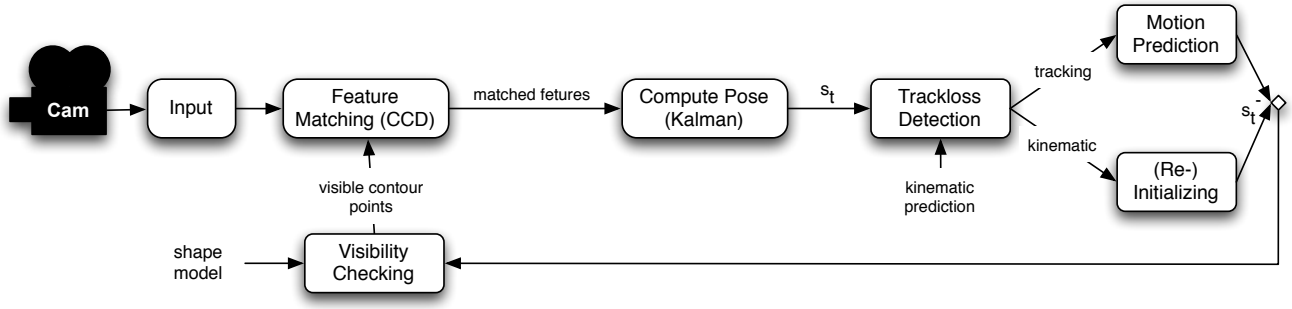


Figure 5. **Tracking Pipeline.** The camera pose can be obtained after calibrating the extrinsic parameters and the overall system. The kinematic measurement of the instrument in 6 degrees of freedom is transfered to a 2D model with 4 DoF (t_x, t_y, h, θ_x). It is used to (re-)initialize the Contracting Curve Density algorithm and to supervise the tracking quality.

C. Tracking with CCD

As already mentioned above, tracking in the context of MIS procedures is exacerbate by changing environment conditions. Simple color segmentation approaches often fail due to varying lightening conditions of different light sources or need a sophisticated fine tuning of parameters. Algorithms that are based upon edge detection suffer from the large amount of feature edges from the background. Figure 7 shows a typically intra-operative scene with an artificial heart and tissue in the background. Neither the Sobel- nor the Canny operator can distinguish the instrument shaft reliable from the background.
The amenity of the CCD modality is that the model's appearance is adjusted over time, since local color statistics are computed in every tracking cycle and maximized according to the shape of the model. Therefore, the method can be applied to marker-based as well as markerless tracking. In fact, the color or texture of the tracked object does not matter, as long as a separation in terms of color between object and background can be achieved. Also a change of the appearance over time (e.g., an account of body liquids) does not disturb the tracking if not the entire object is affected at once.

After setting the initial pose, a Kalman filter generates a prior state hypothesis s_t^- by applying a Brownian motion model to the previous state (s_{t-1}).

$$s_t^- = s_{t-1} + w_t \tag{3}$$

with w being a white Gaussian noise sequence.
The CCD modality requires a sampling of good features for tracking from the object model under the given pose s_t^- and camera view. As a first step, the visible internal and external edges from the polygonal mesh model have to be identified under the current pose hypothesis. Alongside of this contour a set K of uniformly distributed sampling points $\{h_1, \ldots, h_k\}$ is taken to collect color statistics around each sample position on each side of the contour. The basic idea of CCD is to maximize the separation of local color statistics between the two sides of the object boundaries (object vs. background) [21]. The colored shaft of the instrument supports this idea by varying from red tissue and organs. Contemporaneously, the algorithm can account for small change of the shaft appearance over time (e.g., from body liquids), since the statistics are updated in every iteration.
We first sample points along the respective normals, separately collect the statistics, and afterwards blur each statistic with the neighboring ones (cp. Fig. 4). From each contour position h_i, foreground and background color pixels are collected along the normals n_i up to a distance L (that is manually defined and fix), and local statistics up to the 2^{nd} order are estimated

$$\begin{aligned} v_i^{0,B/F} &= \sum_{d=1}^{D} w_{id} \\ v_i^{1,B/F} &= \sum_{d=1}^{D} w_{id} I(h_i \pm L\bar{d}n_i) \\ v_i^{2,B/F} &= \sum_{d=1}^{D} w_{id} I(h_i \pm L\bar{d}n_i) I(h_i \pm L\bar{d}n_i)^T \end{aligned} \tag{4}$$

(5)

with $\bar{d} \equiv d/D$ the normalized contour distance, where the $\pm$ sign is referred to the respective side, and image values I are 3-channel RGB. The local weights w_{id} decay exponentially with the normalized distance, thus giving a higher confidence to observed colors near the contour.
Single line statistics are afterwards *blurred* along the contour, providing statistics distributed on local areas

$$\widetilde{v}_i^{o,B/F} = \sum_j exp(-\lambda\,|i-j|) v_j^{o,B/F}; o = 0, 1, 2 \tag{6}$$

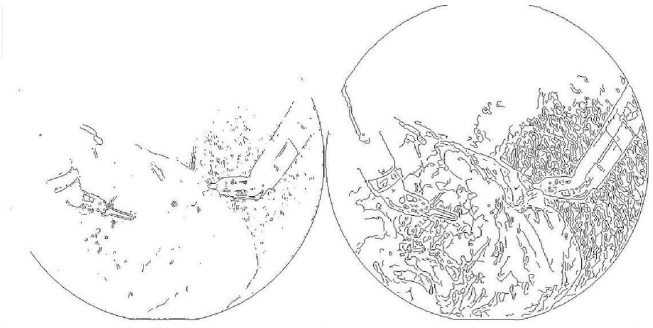


Figure 7. **Edge detection.** The images show edge detection results of the Sobel (left) and the Canny (right) filter. In both cases the tool shaft can hardly be distinguished from background noise. The organ surface comprises many small vessels and structures that raise edges in the vicinity of the tool tip.

and finally normalized

$$\overline{I}_i^{B/F} = \frac{\widetilde{v}_i^{1,B/F}}{\widetilde{v}_i^{0,B/F}} \qquad (7)$$
$$\overline{R}_i^{B/F} = \frac{\widetilde{v}_i^{2,B/F}}{\widetilde{v}_i^{0,B/F}}$$

in order to provide the two-sided, local RGB means $\overline{I}$ and (3×3) covariance matrices $\overline{R}$.

The second step involves computing the residuals and Jacobian matrices for the Gauss-Newton pose update. For this purpose, observed pixel colors $I(h_i + L\overline{d}n_i)$ with $\overline{d} = -1, \ldots, 1$ are classified according to the collected statistics (8), under a fuzzy membership rule $a(x)$ to the foreground region

$$a(\overline{d}) = \frac{1}{2}\left[erf\left(\frac{\overline{d}}{\sqrt{2}\sigma}\right) + 1\right] \qquad (8)$$

which becomes a sharp $\{0; 1\}$ assignment for $\sigma \to 0$; pixel classification is then accomplished by mixing the two statistics accordingly

$$\hat{I}_{id} = a(\overline{d})\overline{I}_i^F + (1 - a(\overline{d}))\overline{I}_i^B \qquad (9)$$
$$\hat{R}_{id} = a(\overline{d})\overline{R}_i^F + (1 - a(\overline{d}))\overline{R}_i^B$$

and color residuals are given by

$$E_{id} = I(h_i + L\overline{d}n_i) - \hat{I}_{id} \qquad (10)$$

with covariances $\hat{R}_{id}$.

Finally the $(3 \times n)$ derivatives of E_{id} can be computed by differentiating (8) and (10) with respect to the pose parameters

$$J_{id} = \frac{\partial \overline{I}_{id}}{\partial s} = \frac{1}{L}\left(\overline{I}_i^F - \overline{I}_i^B\right)\frac{\partial a}{\partial \overline{d}}\left(n_i^T \frac{\partial h_i}{\partial s}\right) \qquad (11)$$

which are stacked together in a global Jacobian matrix $\mathbf{J}_{ccd}$. The state is then updated using a Gauss Newton step:

$$s = s + \Delta s \qquad (12)$$
$$\Delta s = \mathbf{J}_{ccd}^{+}\mathbf{E}_{ccd}$$

The optimization is done until the termination criteria is satisfied $(\Delta s \approx 0)$.

The tracking pose is observed and compared to the kinematic prediction in order to detect tracking loss. This can either be a total loss of tracking, or the sliding of the model alongside of the instrument shaft. For this purpose, we restrict the output of the visual system to lie within a certain range, derived from the current prediction (position and angular values). Since we perform the tracking in a local coordinate frame, we can also easily set pose limits from this values. Furthermore, the estimate of the pose covariance matrix gives a hint for the quality of the tracking. By choosing an empirical maximum threshold for the determinant of the posterior covariance, we can imply a tracking loss.

IV. Application to Visual Guidance

The tracking approach introduced above is utilized to visually guide the surgical instruments and the endoscopic camera.

Although the robotic system is calibrated carefully, the above mentioned inherent imprecisions cannot be determined with a satisfying accuracy to position instruments with a very high precision (which means $1mm$ or below). In particular, the transformation ${}^{F_2}_{I}T$ that follows from the aberration of the carbon fiber shaft of the instrument (cp. Fig. 3(c)) cannot be minimized to a satisfying amount.

Visual servoing is a popular approach to guide a robotic appendage (i.e., a surgical instrument in our case) using visual feedback from a camera system. In general, visual servoing can roughly be divided into two categories: position-based visual servoing control (PBVS), in which a Cartesian coordinate is estimated from image measurements and image-based visual servoing (IBVS) approaches, which seek to extract features directly from an image series. In general, the accuracy of image-based methods for static positioning tasks is less sensitive to calibration than PBVS [25]. Image-based servoing does not depend as much on calibration as the error is reduced directly in image pixels. However, a practical difficulty during the alignment of surgical instruments with a desired position in space lies in the fact that the instrument is not necessarily in the field of view of the camera and therewith no image-features can be extracted. Hence, we first drive the instrument to a Cartesian coordinate (reconstructed using stereopsis of the 3D endoscope) which is in the field of view of the camera.

Since the 3D reconstruction suffers from a certain error (caused by the mentioned intrinsic errors) we continue with image-based servoing to overcome the remaining distance. In fact, an eligible point close to the final position is chosen.

Given a target position that the robot is to reach, visual servoing aims to minimize an error $e(t)$, typically defined by

$$e(t) = s(m(t), a) - s^* \tag{13}$$

where s^* represents the target pose, $s(m(t), a)$ the measured pose, $m(t)$ the measured image feature points and a any additional knowledge needed, such as information from the camera calibration. The function $s(m(t), a)$ characterizes the end point of the tool tip of an instrument carried by the robot. In PBVS the position of the tracked features is extracted from the camera image coordinates and projected to the world frame by the mapping a, determined during camera calibration. The target position can be extracted from image features in a similar way. While PBVS minimizes the error $e(t)$ in world coordinates and the camera is treated as a 3D positioning sensor, IBVS directly tries to find a mapping from the error function to a commanded robot motion.
As mentioned above, PBVS is used to drive the instrument to a reconstructed point, which is located within the view of the camera. As soon as this point is reached, the remaining distance to the target goal is minimized in image coordinates. In many IBVS scenarios the camera is attached to the robot which is to be commanded (eye-in-hand configuration) and therewith the velocity of the camera ξ is calculated. In our setup, the instrument and the endoscope are carried by two different robots and the calculated velocity ξ has to be transformed to the robot that carries the instrument.
A single image feature, for instance the tip of an instrument or a carried needle, is tracked in both left and right camera coordinates. The feature vector $s = (x_L, x_R)^T = (u_L, v_L, u_R, v_R)^T$ comprises these coordinates:

$$s(t) = \begin{bmatrix} u(t) \\ v(t) \end{bmatrix} \tag{14}$$

Its derivative $\dot{s}(t)$ is referred to as image feature velocity. It is linearly related to the camera velocity $\xi = [\ \upsilon \quad \omega\ ]^T$, which is composed of linear velocity υ and angular velocity ω. The relationship between the time variation of the feature vector s and the velocity in Cartesian coordinates ξ is then established by

$$\dot{s} = L_s \xi \tag{15}$$

where L is the *interaction matrix* or *image Jacobian* [26]. The interaction matrix L_x related to an image point $x = (u, v)^T$ reads as follows:

$$L_x = \begin{bmatrix} -\frac{1}{z} & 0 & \frac{u}{z} & uv & -(1+u^2) & v \\ 0 & -\frac{1}{z} & \frac{v}{z} & 1+v^2 & -uv & -u \end{bmatrix} \tag{16}$$

Variable z represents the depth of a point relative to the camera frame. There exist different ways to approximate the value of z, for example via triangulation in a stereo setup or via pose estimation. Most of the existing methods assume an calibrated camera, even if the impact of the calibration is not very high. Few systems even assume a constant depth of the tracked feature and therewith a constant image Jacobian. In our approach, variable z is estimated via the kinematic chain of the system. The interaction matrix can then be updated online and the approach is easily transferable to miscellaneous camera configurations. For instance, we equipped another robot arm with a second monocular Fujinon™endoscope that provides a different view on an object. Using equations (13) and (15) we obtain $\dot{e} = L_e \xi$ and our final control law

$$\xi = \lambda L_e^+ e \tag{17}$$

where λ is a positive gain factor and L_e^+ the Moore-Penrose pseudo-inverse of L_e.
As mentioned above, a single visual feature s is tracked in the left and right images equation (15) is rewritten as

$$\begin{bmatrix} \dot{x}_L \\ \dot{x}_R \end{bmatrix} = \begin{bmatrix} L_L \\ L_R \, {}^R_L V \end{bmatrix} \xi_L \tag{18}$$

The spatial motion transform ${}^R_L V$ to transform velocities expressed in the right camera frame R to the left camera frame L is given by

$${}^R_L V = \begin{bmatrix} {}^R_L R & S(t)\, {}^R_L R \\ 0 & {}^R_L R \end{bmatrix} \tag{19}$$

where $S(t)$ is the skew symmetric matrix associated with the linear transformation vector t and where (R, t) is the transformation from the left to the right camera frame.

To consider the characteristics of the trocar kinematic during minimally invasive surgery, the instrument movement at the fulcrum has to be zero in all directions that are perpendicular to the instrument shaft. Since the location of the incision point is well-known from the simulation software, the 6DoF motion of the robot can be constrained to 4DoF at the trocar. The velocities ${}^T_T \xi = ({}^T_T \upsilon, {}^T_T \omega)^T$ at the trocar point T and the velocities ${}^I_I \xi = ({}^I_I \upsilon, {}^I_I \omega)^T$ of the instruments tip I are related as follows:

$$\begin{aligned} {}^T_I V \cdot {}^I_I \xi &= {}^T_T \xi \\ \Leftrightarrow \begin{bmatrix} {}^T_I R & S({}^T_I t)\, {}^T_I R \\ 0 & {}^T_I R \end{bmatrix} \begin{bmatrix} {}^I_I \upsilon \\ {}^I_I \omega \end{bmatrix} &= \begin{bmatrix} {}^T_T \upsilon \\ {}^T_T \omega \end{bmatrix} \end{aligned} \tag{20}$$

Assuming a straight shaft, ${}^T_I R$ is the identity matrix and $t = (0, 0, d)^T$ with d being the insertion depth of the instrument. Since only the z-direction (the direction of the shaft) is free to move, the linear velocity at the insertion point is denoted by ${}^I_I \upsilon = (0, 0, {}^I_I \upsilon_z)$. Solving (20) yields to

$${}^I_I \omega_x = -\frac{{}^I_I \upsilon_y}{d} \quad \text{and} \quad {}^I_I \omega_y = \frac{{}^I_I \upsilon_x}{d} \tag{21}$$

So far, we covered the control of surgical instruments. Furthermore, automated camera control (e.g., the endoscope automatically follows an instrument) is also of high interest to assist surgeons. The control law is similar to the instrument control, but in contrast, we prohibit movements in the directions of the shaft. For safety reasons of the patient it is not suitable that the endoscope induces depth motion. Taking Eqn. (21) into account and setting the camera velocities ${}^C_C v_z = {}^C_C \omega_z = 0$ we obtain the new interaction matrix L_{cam}

$$\begin{aligned} \dot{s} &= \begin{bmatrix} L_v \\ L_\omega \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix} \qquad (22) \\ &= \begin{bmatrix} -\frac{1}{z} & 0 \\ 0 & \frac{1}{z} \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} \\ &+ \begin{bmatrix} xy & -(1+x^2) \\ 1+y^2 & -xy \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} -\frac{1}{z} - \frac{1}{d}(1+x^2) & -\frac{1}{d}xy \\ -\frac{1}{d}xy & -\frac{1}{z} - \frac{1}{d}(1+y^2) \end{bmatrix}}_{L_{cam}} \begin{bmatrix} v_x \\ v_y \end{bmatrix} \end{aligned}$$

V. Experimental Results

In order to evaluate the performance of the tracking system, more or less crucial instrument poses during system operation have been taken. After presenting the results of the tracking, the compliance of the trocar during visual guidance of an instrument is shown.

The evaluation has been performed on a Intel Xeon QuadCore™2.4Ghz system. Images were taken and processed in real-time with full PAL resolution (768×576) from the framegrabber. As a first step, the precision of the instrument projection into image space, derived from the kinematic data, was verified (cf. Fig. 12, first image). The

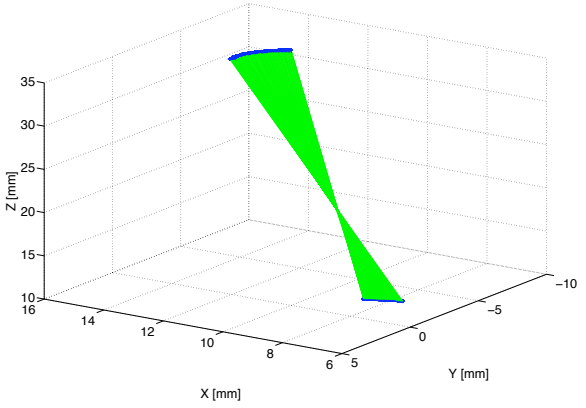


Figure 11. Evaluation of the trocar constraint by means of a magnetic tracking system

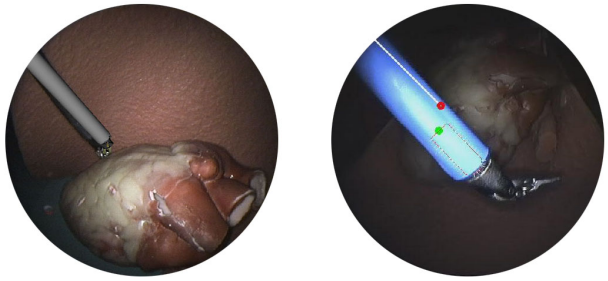


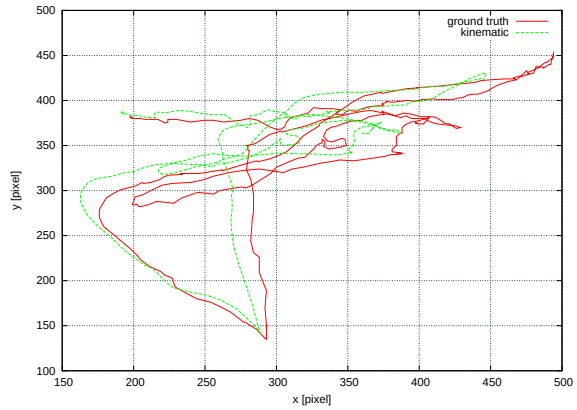
Figure 12. Left image: Initial estimation of the instrument pose, derived from sensor readings. Right image: Mismatch of the model scaling factor and the instrument due to strong specular reflections.

data is received by the tracking framework via network and applied to the CAD model of the instrument. To project the instrument pose into image space, a virtual camera is set up in a similar fashion, with position and orientation equal to the real endoscope. The projection of the shaft does not have to overlay the instrument that is to be tracked perfectly, but a good match supports a fast initialization of the tracking. A good agreement of projection and instrument helps to keep the normals of the sampled contour points smaller, making the tracking more robust and faster. The search length was determined experimentally.

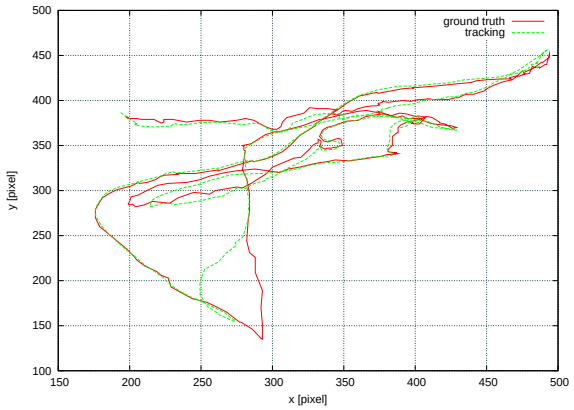
For evaluation purposes, a series of images was annotated by hand. Figure 8(a) shows the projection of the kinematic prediction into image space in comparison with the ground truth. The average distance error calculated over all frames is around 33 pixels. The offset between the estimation and the actual position of the instrument is well observable. The average distance between the tracked point and the ground truth could be reduced to 5.7 pixels. For the depicted image series, tracking was lost one time for a period of approximately 10 frames (lower left side in Fig. 8(b)). The plots of the tracking $x-$ and $y-$ errors (Fig. 8(c) and Fig. 8(d)) point out a fast reinitialization of the tracking around frame number 170. Excluding the 10 frames of the tracking loss, the accuracy can further be reduced to $4.6px$.

As already stated, the presented approach is not limited to a specific appearance of the instrument. In fact, we used three different instruments with blue, red and gray colored shafts. In our artificial environment, the background is very dark, since no brightened ribcage or abdominal wall limits the sight. Therewith, tracking the gray shaft is most challenging. While the detection of the shaft itself works flawless, more sliding of the model is observable, compared to the blue or red shaft. Since the color of the distal end of the instrument changes from gray to silver, no hard contour is given anymore. In this case, CCD loses tracking during fast movements.

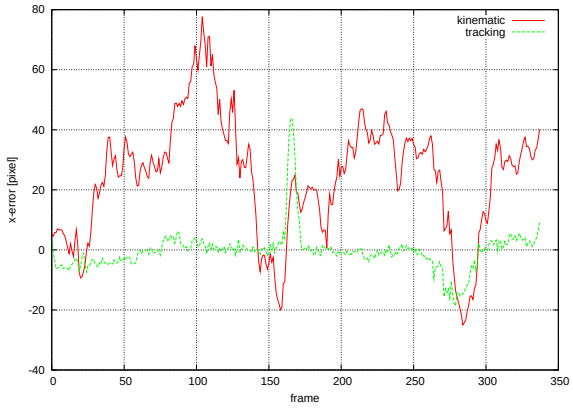
The main flaw of the proposed approach is the detection of the accurate scaling factor. Since the CCD algorithm seeks to maximize color statistics alongside of the contour



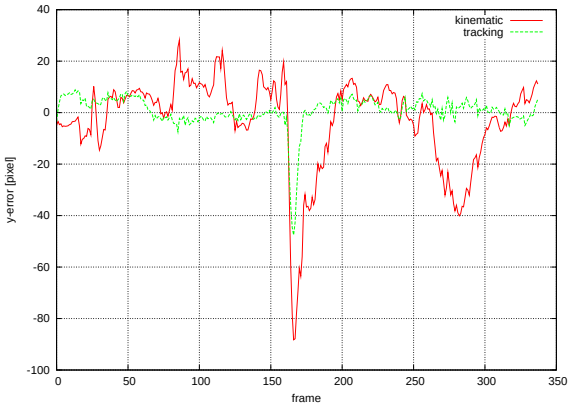
(a) Kinematic Prediction vs. ground truth in image space



(b) Tracking vs. ground truth in image space



(c) Errors compared to ground truth (x-axis)



(d) Errors compared to ground truth (y-axis)

Figure 8. **Tracking errors:** The ground truth data was annotated by hand. Figures 8(a) and Fig. 8(b) show the error of the kinematic prediction vs. the ground truth and the tracking in image space respectively. Figure 8(c) and Fig. 8(d) depict the errors in $x-$ and $y-$ direction in pixels, compared to the ground truth.

edges between model and background, strong specularities at the shaft can distort the measurement and are spuriously recognized as part of the instrument (cp. Fig. 12). Those kind of reflections especially appear at low distances (less than $3cm$) between the instrument and the light source, dependent on the present luminosity. Then, the center of the shaft can still be located accurately, but the distance to the tip is wrong.

Regarding the visual guidance, we evaluated the compliance with the trocar point, in addition to the experiments that have been performed in the original work. Therefore, we utilized a magnetic tracking system (Polhemus Liberty^{TM}). Since the magnetic markers were attached at the distal end of the instrument, the influence of the robot motors can be neglected. Figure 11 shows the movement of the instrument and the fulcrum.

VI. Conclusion and Future Works

This paper has explored the tracking of surgical instruments in minimally invasive surgery and its application to the visual guidance of the instruments and the endoscopic camera. Encoder readings from the robots were used to predict an approximated pose of the instrument in image space. The approximation is then used to (re-)initialize the image-based tracking, to set pose limits and to supervise a tracking loss. As modality, the Contracting Curve Density algorithm was used, which maximizes local color statistics collected at the model contour in order to separate it from the background.

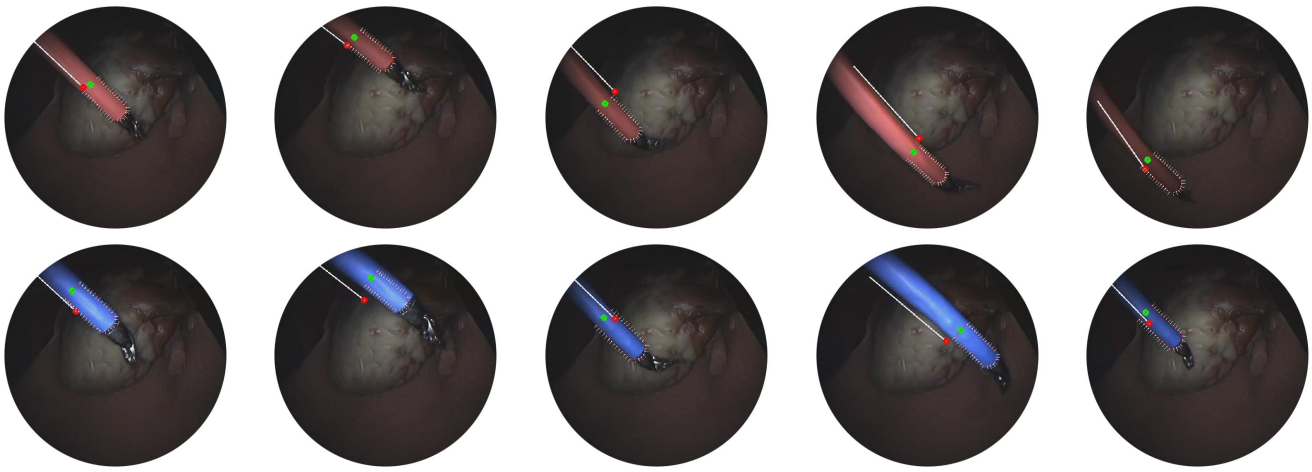


Figure 9. Top row: Tracking a reddish instrument. The tracking is very stable, even is the shaft color is similar to the background. The white line (ending in a red dot) indicates the projection of the kinematic prediction, the green dot is dedicated to the actually tracked position. The images are overlie with the instrument model and the contour normals used for sampling the local color statistics. Bottom row: Instrument tracking with a blue shaft. Since CCD does not employ a color or texture map of the instrument it can be applied to various shaft colors without changes.

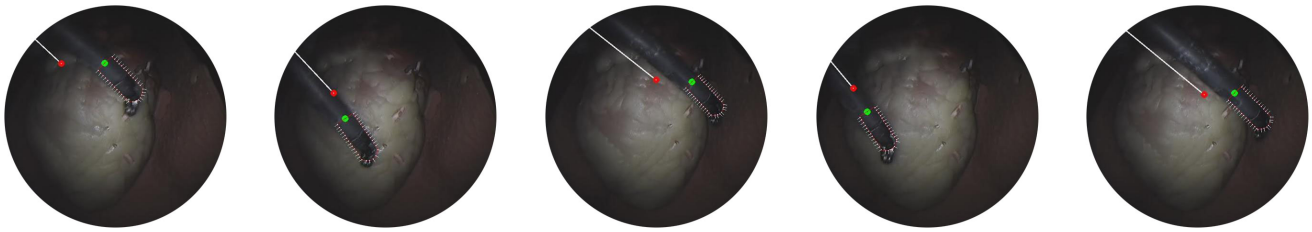


Figure 10. Tracking a grayish instrument shaft. As background, artificial skin and a heart model was used. In a real laparoscopic intervention the depth field would be more restricted, resulting in a brighter and more uniform illumination of the scene.

The performed experiments are very promising. Without the need for changing the model or program parameters, a blueish and a reddish instrument was tracked accurately. Problems in finding an adequate scaling factor can arise due to specular reflections, if the distance between instrument and light source is small. During a preprocessing step, this reflections could be removed, since their location and pixel values are well-known. In order to prevent a misplacement of the model at the distal end of the shaft, a non-uniform distribution of sampling normals could be introduced. After splitting up the model into an articulated model with two parts, one representing the shaft and one representing the rounded tip, the number of sampling points at each sub-model could be adjusted independently. As the statistics that are collected at the shaft would be weighted more than the statistics at the end, a shift could presumably be prevented. Also a simple blob detection that looks for the silver-colored forceps could be employed in addition.

ACKNOWLEDGMENT

This work is supported by the German Research Foundation (DFG) within the Collaborative Research Center SFB 453 on “High-Fidelity Telepresence and Teleaction”.

REFERENCES

[1] C. Staub, A. Knoll, T. Osa, and R. Bauernschmitt, “Autonomous high precision positioning of surgical instruments in robot-assisted minimally invasive surgery under visual guidance,” in *Proceedings of the IEEE International Conference on Autonomic and Autonomous Systems*, Cancun, Mexico, March 2010, pp. 64–69.

[2] G. Guthart and J. Salisbury, J.K., “The intuitive^{TM}telesurgery system: overview and application,” in *Proceesings of the IEEE International Conference on Robotics and Automation*, vol. 1, 2000, pp. 618–621.

[3] H. Mayer, D. Burschka, A. Knoll, E. Braun, R. Lange, and R. Bauernschmitt, “Human-machine skill transfer extended by a scaffolding framework,” in *Proceesings of the IEEE International Conference on Robotics and Automation*, may 2008, pp. 2866 –2871.

[4] H. Mayer, I. Nagy, A. Knoll, E. Braun, R. Lange, and R. Bauernschmitt, "Adaptive control for human-robot skill-transfer: Trajectory planning based on fluid dynamics," in *Proceesings of the IEEE International Conference on Robotics and Automation*, april 2007, pp. 1800 –1807.

[5] H. Wakamatsu, A. Tsumaya, E. Arai, and S. Hirai, "Manipulation planning for knotting/unknotting and tightly tying of deformable linear objects," in *Proceedings of the IEEE International Conference on Robotics and Automation*, April 2005, pp. 2505–2510.

[6] C. Staub, T. Osa, A. Knoll, and R. Bauernschmitt, "Automation of tissue piercing using circular needles and vision guidance for computer aided laparoscopic surgery," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2010, pp. 4585–4590.

[7] F. Nageotte, P. Zanne, C. Doignon, and M. de Mathelin, "Stitching planning in laparoscopic surgery: Towards robot-assisted suturing," *The International Journal of Robotics Research*, pp. 1303–1321, 2009.

[8] P. Hynes, G. Dodds, and A. Wilkinson, "Uncalibrated visual-servoing of a dual-arm robot for mis suturing," in *Proceedings of the IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics*, Feb. 2006, pp. 420–425.

[9] A. Krupa, J. Gangloff, C. Doignon, M. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux, "Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 842–853, Oct. 2003.

[10] G.-Q. Wei, K. Arbter, and G. Hirzinger, "Real-time visual servoing for laparoscopic surgery. controlling robot motion with color image segmentation," *IEEE Engineering in Medicine and Biology Magazine*, vol. 16, no. 1, pp. 40–45, Jan.-Feb. 1997.

[11] A. Casals, J. Amat, and E. Laporte, "Automatic guidance of an assistant robot in laparoscopic surgery," in *Proceesings of the IEEE International Conference on Robotics and Automation*, vol. 1, Apr 1996, pp. 895–900.

[12] C. E. Reiley and G. D. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in *Proceedings of the International Conference on Medical Imaging and Computer Assisted Interventions*, 2009, pp. 435–442.

[13] H. C. Lin, I. S. ans David Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, September 2006.

[14] D. R. Uecker, C. Lee, Y. F. Wang, and Y. Wang, "Automated instrument tracking in robotically-assisted laparoscopic surgery," *Journal of Image Guided Surgery*, vol. 1, pp. 308–325, 1998.

[15] C. Doignon, F. Nageotte, and M. de Mathelin, "The role of insertion points in the detection and positioning of instruments in laparoscopy for robotic tasks," in *Proceedings of the International Conference on Medical Imaging and Computer Assisted Interventions*, 2006, pp. 527–534.

[16] S. Voros, J.-A. Long, and P. Cinquin, "Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders," *The International Journal of Robotics Research*, vol. 26, no. 11-12, pp. 1173–1190, 2007.

[17] D. Burschka, J. J. Corso, M. Dewan, W. W. Lau, M. Li, H. C. Lin, P. Marayong, N. A. Ramey, G. D. Hager, B. Hoffman, D. Larkin, and C. J. Hasser, "Navigating inner space: 3-d assistance for minimally invasive surgery," *IEEE Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 5–26, 2005.

[18] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2009, pp. 3940–3947.

[19] Y. Shiu and S. Ahmad, "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form ax=xb," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 1, pp. 16–29, Feb 1989.

[20] G. Welch and G. Bishop, "An introduction to the kalman filter," University of North Carolina at Chapel Hill, Department of Computer Science, Tech. Rep., 2006.

[21] G. Panin, A. Ladikos, and A. Knoll, "An efficient and robust real-time contour tracking system," *Computer Vision Systems, International Conference on*, vol. 0, p. 44, 2006.

[22] G. Panin, E. Roth, and A. Knoll, "Robust contour-based object tracking integrating color and edge likelihoods," in *VMV*, 2008, pp. 227–234.

[23] A. Ruf, M. Tonko, R. Horaud, and H.-H. Nagel, "Visual tracking of an end-effector by adaptive kinematic prediction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, Sep 1997, pp. 893–899 vol.2.

[24] L. Itti and C. Koch, "Computational modelling of visual attention." *Nature reviews. Neuroscience*, vol. 2, no. 3, pp. 194–203, March 2001.

[25] B. Espiau, "Effect of camera calibration errors on visual servoing in robotics," in *The 3rd International Symposium on Experimental Robotics*. London, UK: Springer-Verlag, 1994, pp. 182–192.

[26] F. Chaumette and S. Hutchinson, "Visual servo control. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, Dec. 2006.

Applying the Theory of Constraints to Health Technology Assessment

Johan Groop, Karita Reijonsaari, and Paul Lillrank
Institute of Healthcare Engineering, Management and Architecture (HEMA)
Aalto University
Espoo, Finland
johan.groop(a)tkk.fi, karita.reijonsaari(a)tkk.fi, paul.lillrank(a)tkk.fi

***Abstract*—Applying assessment methods commonly used in healthcare, such as randomized, controlled trials, and financial analysis, often proves challenging when assessing technologies that strive to improve productivity. The traditional way of seeking the procurement and implementation of process-improving technology is based on financial analyses. These are often not only laborious, but also based on weak assumptions about the interrelationship between technology, processes, and the institutional environment. There is a need for a more pragmatic managerial approach. This paper suggests an alternative means, based on the theory of constraints (TOC), arguing that the primary focus of technology assessment should be on the ability of technologies to remove or alleviate organizational constraints, in order to increase throughput rather than reduce costs. Ultimately, the latter will happen as a consequence of the former. The suggested approach is illustrated through a case study of a home care organization, in which improved productivity is sought through the implementation of a mobile solution. Although conventional financial reasoning held that the implementation would save time and therefore be beneficial, the TOC approach showed that the implementation would in fact have an adverse effect under the current mode of operation.**

Keywords*—*Theory of Constraints, technology assessment, home care, healthcare operations management, mobile technology

I. INTRODUCTION

Most healthcare providers are urged to improve their productivity to cope with increasing demand under resource constraints. There is a rising trend of seeking improvements in productivity through the use of technology, ICT (information and communication technology) in particular. This paper does not discuss clinical technologies, since their assessment follows a different logic and their methodologies are well developed. Instead the paper focuses on technologies that seek to improve productivity through process improvements (henceforth process technologies), following an OM (Operations Management) perspective.

The number of different process technologies on the market is accumulating fast. Decision makers would like to base their decisions on hard numbers, such as net present value (NPV), cost-benefit, and payback time. Data allowing such analyses are, however, rarely available. Evaluating the financial measures of health technologies properly is difficult, for the following reasons [1]:

- Technology solutions often impact on processes and may alter a service as a whole. Therefore it is difficult to evaluate the influence in advance, particularly if the operational mechanisms of a service are not fully understood. There may be opportunity costs and system effects, such as an improvement in one part of a process being offset by adverse impacts on other parts.
- The gold standard of intervention assessment is the randomized, controlled trial (RCT), commonly used in clinical research (e.g., [2]). RCT requires detailed methodological design, including a statistically significant sample and a coherent control population. It can only be done for one technology at a time; once the technology has been implemented. RCT is often too time-consuming and expensive for fast-paced process technology evaluation.

There is a need for a more pragmatic, managerial approach to evaluating process-improving technologies. Such an approach would need to focus on the ability of a particular technology to move the organization towards its goal. Every organization has a goal and defining it clearly is of great importance. The generic goal of any for-profit business organization is to "make money now as well as in the future" [3], "without violating the necessary conditions of providing a satisfying work environment for employees and ensuring customer satisfaction" [4], e.g., by offering quality products or services. The two prerequisites should not be confused as separate goals. They are threshold conditions which need to be satisfied at least to some minimum level, above which their impact on performance diminishes. The goal, on the other hand, has no upper limit, and it is something that should constantly be pursued.

In public healthcare, defining the goal is more ambiguous, because of the inherent complexities. Several authors have suggested defining the goal of healthcare in terms of different outcome-related measurements [5][6][7][8][9], while others have chosen to define the goal in terms of tangibles, such as money [10][11], and the volume of services produced [12] or patients treated [13]. There is no obvious correlation between the volume of procedures and actual health outcomes; sometimes less is more [14]. However, the question of which volume and combination of treatments leads to the optimal outcome is beyond the scope of OM. Therefore efficiency studies must be based on the assumption that a certain clinically justified

level of service of a given quality is necessary, and the challenge is to produce those at the lowest (or optimal) cost.

Taking on an OM perspective, Wright and King [13] build on the generic goal, suggesting that the goal of public healthcare systems is to "treat more patients, better, sooner, now and in the future". The authors choose the definition of Ronen et al. [12], who propose that the highest-level goal of a not-for-profit organization, such as a publicly financed health organization, is to "maximize quality healthcare services provided to its customers, subject to budgetary constraints". Both definitions assume that the two necessary conditions are satisfied. The condition of satisfying customers presumes that the right level and quality of service is provided to patients who need it. The authors also find that both definitions implicitly incorporate the pursuit of better health outcomes.

In order to measure an organization's performance relative to its goal, the goal needs to be translated into operational language through clear, simple and appropriate performance measures [15]. Thus, the goal should be defined as something easily measurable [3]. Financial measures, such as net-profit or ROI (return-on-investment), are affected by operational performance measures; therefore focusing on cost measures without a proper analysis of operational indicators is misleading. A certain technology can promote the operational performance measures derived from the organizational goal.

Organizations should focus on throughput, defined as the rate at which the organization achieves its goal [3]. In a for-profit organization where the goal is to make money, throughput is defined as "the rate at which a system generates money through sales" [3]. In not-for-profit organizations defining throughput is more complex and varies according to the specific characteristics of the organization. For instance, in elective surgery throughput can be defined as procedures performed (output) minus rework. Arguing that "attainment of the financial goal [i.e., making enough money to cover expenses] provides for the realization of the clinical goal" (i.e., high-quality care), Gupta and Kline [10] define throughput as the income generated from patient care.

Throughput is not the same as output, although the concepts are related. In manufacturing a finished product is considered output, which is turned into throughput once it is sold [3]. As services are produced and consumed simultaneously, the distinction between output and throughput is less obvious, particularly if payment is made prior to production and consumption. The practical definition of throughput depends on how we define the goal, e.g., money earned, patients treated, or service time produced. As a crucial operational indicator, it is important that throughput be defined as something manageable by a producer, serving its purpose, and that it is easily measurable. In public healthcare, throughput is therefore often defined as something tangible, such as the number of patients treated [13], services rendered [12], or money generated through patient care [10].

This paper takes a healthcare operations management approach, focusing on the managerial aspects of health technology assessment (HTA). The framework follows the management philosophy of the theory of constraints (TOC), in that it focuses on the impact of technologies on organizational constraints. A constraint is generally defined as "anything that limits a system from achieving higher performance versus its goal" [3]. By exploiting or breaking a system's constraints, performance can be improved. The suggested approach explains how constraints management can be used as a tool for technology assessment.

Section 2 provides an introduction to the theory of constraints. The objective is to clarify the logic behind TOC, and to provide reasoning about why it would be a suitable tool for HTA. First, the TOC philosophy and its main principles are explained, and different kinds of constraints presented. The more traditional strategy of reducing operational expense is contrasted with the TOC approach to maximizing throughput. Section 3 describes how TOC can be used as a tool for technology assessment. Section 4 illustrates this approach using an example from an ongoing case study of a home care unit, regarding a decision as to whether or not to invest in a certain mobile platform-based technology solution. In Section 5, we discuss the limitations of the study, the suitability of the TOC approach, and how it relates to financial analysis. Section 7 provides a conclusion and presents suggestions for further research.

II. THEORY OF CONSTRAINTS

Originally developed by Eliyahu M. Goldratt [15], "TOC views every organization as a chain of interdependent events (or processes) where the performance of each event (or process) is dependent upon the previous event" [5]. TOC is based on the assumption that every organization has at least one (but no more than a few) constraint(s) that keeps it from reaching a higher level of performance. Without a constraint the system's performance would be infinite [16]. Although there is a tendency to think of constraints in physical terms, e.g., a lack of hospital beds, inadequate MRI capacity, or a shortfall in staff, it has been shown that the majority of constraints are not physical but policy constraints, such as operational procedure or management policy [11], which in turn may cause resource constraints. TOC argues that the system can only be improved by strengthening its weakest link, i.e., the system's constraint. While an organization may experience several difficulties, some problem (constraint) has to be the most significant for the organization's ability to reach its goal [4]. Therefore any improvement effort should target the system's constraints. In this respect TOC differs from other management philosophies, such as Total Quality Management (TQM) and Just-In-Time (JIT), which consider any improvement in a process as being an improvement of the system as a whole [17].

Figure 1. A three-step process and its capacity and constraints (adapted from Ronen et al. [12])

Figure 1 illustrates the logic of TOC through a simplified example. It shows a three-stage process and the capacity of each stage, as well as its demand and output. The demand on the system is for 100 units (e.g., the treatment of 100 patients). The first department can handle 100 units per day, the second department 50 units, and the final department 75 units per day. Department 2 constitutes the system constraint because it can only process 50 patients per day. The system throughput, or in this case the system output, can only be increased by improving the performance of the constraint [3]. "An hour lost at the constraint is an hour lost to the whole system" [15]. Departments 1 and 3 have an overcapacity of 50 and 25 units respectively per day. Fully utilizing the capacity of either department will not impact on the output of the system, but will build up patient-in-process (PIP) inventory [18] (PIP is the healthcare equivalent to work-in-process in manufacturing). The PIP will lengthen the time it takes for patients to get through the system (lead time), as well as increase the ability of the system to respond to new or altered medical needs (response time). This has a negative effect on patient satisfaction. It may also affect clinical quality and thereby increase the operating expenses of the system. Rather than aiming for greater efficiencies through full resource utilization of departments 1 and 3, the system should aim to keep the production flow at a steady rate (determined by the constraint), by ensuring that department 2 can process its full capacity (50 patients per day).

TOC distinguishes between constraints and everything else, which is referred to as non-constraints. According to the above-mentioned logic, the ability of the system to produce is not improved by increasing the efficiency of non-constraints.

A. The Five Focusing Steps of TOC

Much like the other OM philosophies, TOC includes a process of continuous improvement, incorporating five focusing steps [3][19]: "1) *identify* the constraint(s); 2) *exploit* the constraint(s); 3) *subordinate* everything else to the decision taken in step 2; 4) *elevate* the constraint(s), and 5) if a constraint is *broken*, go back to step 1. Do not let *inertia* be the cause of a constraint." Step 2 refers to making sure the constraint is fully utilized at all times (provided it is a resource constraint), and eliminating policy constraints. Step 3 means that the constraint determines the pace of production. The main priority of the non-constraints is to keep the constraint(s) fully utilized at all times, their own utilization being of much less significance. Step 4 means improving the performance of the constraint(s) (e.g., by increasing its capacity, eliminating unnecessary work etc.). Step 5 expresses the need to repeat the process.

B. The Need for a Stable System

A widely held assertion of OM is that a swift even production flow increases performance [20][21]. Other widespread OM philosophies such as TQM, Lean, and Six Sigma seek to balance the production flow by *balancing the capacity of the system*, so that every step produces at the same rate [22]. Efficiency is pursued though waste reduction; the capacity of each step should equal demand (e.g., a capacity and production rate of 50 units per department in the above-mentioned example). An even flow is sought by emphasizing the reduction of both internal and external fluctuations, in terms of poor quality, the processing time of each production step, quality, worker absence, lateness etc.

TOC, on the other hand, recognizes that fluctuations can never be completely removed. This is particularly true in open systems, such as healthcare, which includes considerable customer-induced variability [23]. Because of the inevitable existence of fluctuations, TOC argues that a balanced system ultimately becomes inefficient. Fluctuations at steps that have no protective capacity cause inventory to build up, prolonging lead times, and reducing responsiveness etc. [24]. The fluctuations are caused by the combination of two phenomena: 1) internal fluctuations, i.e., the statistically inevitable fluctuations inherent in each step, and 2) cumulative fluctuation, caused by dependent events, i.e., the existence of a process. The fluctuations of each step accumulate, rather than level out, causing greater fluctuations downstream [15].

TOC strives for an even production flow, but by means of an *unbalanced system*. This refers to the existence of at least one constraint. Contrary to the conventional view of constraints – something negative that should be eliminated – TOC regards constraints as pace setters, around which the rest of the system should be managed, or *subordinated*. The basic premise is that, by managing the whole system according to the production rate of the constraint, a stable flow can be reached (as opposed to by balancing the capacity and production rate).

According to TOC, before the performance of a system can be improved it must first be stable. This requires the constraint to be identified, so that the non-constrained steps can be subordinated to the constraint, creating an unbalanced stable system. If no constraint exists, TOC suggests creating one, preferably one that is as natural as possible, e.g., an expensive piece of equipment. The positioning of the constraint is an important strategic decision [25].

C. Traditional Focus on Cost

Many healthcare organizations tend to follow a cost reduction policy, seeking to reduce all costs while simultaneously improving the efficiency of resources [5]. Although this is an intuitively logical approach, practice has shown that the measures, decisions, and behaviors this approach leads to can have quite the opposite effect [3][12][15][11]. If greater efficiency of all the steps in a process, instead of the constraints, is sought, PIP builds up, leading to longer waiting times, lead times, and response times, which in turn increases operating expenses. Traditional cost accounting creates incentives that encourage high efficiencies [3]. One reason for is that it is based on the false assumption that all resources are fully utilized at all times [15].

Cost accounting was first developed to deal with the accounting complexities stemming from the rise of mass production in the 1920s. At the time "direct wages were a major component in costs and were considered real variable costs […] The indirect costs were low and assigning them to products based on real direct wages (or another volume-based variable such as machine hours) was a reasonable approximation for making business decisions such as continued production of a product, investment decisions, buy-or-make decisions, and so on" [12]. In those days, indirect costs (i.e., overhead) constituted around 5-10 percent of the total costs of production. Today the indirect costs are 20-80 percent. Assigning them to products or services on the basis of the cost accounting formula creates severe distortions [3][12][15]. For instance, it fails to distinguish between constraints and non-constraints, creating incentives to improve the efficiency of non-constraints, which will not increase the throughput of the system as a whole. In fact, it may cause throughput to decrease, as the flow becomes unstable. This is also known as "the efficiencies syndrome" [12]; a situation where increasing the efficiency of every production step, "emphasizing the utilization of inputs instead of focusing on outputs", reduces the efficiency of the whole system.

D. Focus on Throughput

TOC holds that the most important performance measure of an organization is its performance relative to its goal, i.e., the throughput in terms of the number of "units of the goal" [5]. To improve performance, TOC focuses on maximizing throughput rather than reducing operational expense. This is also known as "throughput orientation" [4][26] or "throughput-world thinking" [26]. The logic is based on the assumption that the change in operational thinking will lead to behavior and decision making that reduces costs, through reduced lead time, response time, and PIP, as well as improved service quality [12]. Operating expenses can only be reduced to a certain level, while throughput, in theory, can be increased infinitely [3].

As the production rate of the primary constraint determines the throughput of the system, the cost of the constraint is the cost of the entire system divided by the available number of constraint hours [15]. This approach is also referred to as throughput accounting (TA) [3][27]. It reduces the complexity by focusing on the production capacity and cost of the production step that matters, i.e., the constraint, instead of all steps separately.

III. An Alternative Approach To Health Technology ASSESSMENT

When health technologies are being assessed, the focus should first be on the technologies' impact on throughput, then on costs. When estimating the time-saving effect of a new technology, it is tempting to evaluate the saved cost in terms of cost per hour. If the time was saved at a constraint this approach might be valid. However, if the time saved does not affect the capacity of the constraint, it is unlikely to increase throughput [1].

Most healthcare organizations have large fixed costs, mainly the cost of personnel. Typically, only 10-15 percent of the budget is variable costs [3]. Saving the time of non-constrained labor (mainly a fixed cost) does not affect unit costs, unless the time saved can be allocated elsewhere (e.g., moving excess capacity to the constraint to increase throughput) or less labor is required as a consequence, while still maitaining full utilization of the constraint. Unit costs, on the other hand, can be reduced by improving the throughput of the organization, as this spreads the fixed costs of the organization over more produced units.

By concentrating on assessing the impact health technologies have on the system's constraints, and thereby on throughput, it is possible to avoid some of the complexities associated with technology assessment in healthcare. If throughput can be increased at a reasonable cost, the additional expense is likely to be more than offset by the increase in throughput.

The first step is to identify the system's constraints using the five focusing steps [3]. Second, the effect of the alternative technologies on the constraints is evaluated. Which constraints does a certain technology affect and how does it improve throughput? Will the technology remove or alleviate the constraint to allow increased throughput? Does the implementation of a certain technology, and the accompanying process changes it makes possible, affect the constraint directly, or only in combination with some other technology or improvement effort?

If it is determined that a technology can remove or alleviate a resource constraint (a policy constraint is not likely to be affected by technology), the marginal return on an investment in the constraint should be examined [25]. In other words, how much additional throughput can the investment achieve? Here, again, focusing on the constraint can reduce complexity.

When assessing different technologies it is not uncommon to encounter a situation where there are several

constraints, as well as alternative technological solutions that may affect more than one constraint. Such a situation is illustrated in the following case study.

IV. THE STUDY

This section descibes the TOC approach in use through an empirical example. The investigation illustrated in this paper was performed as part of a larger ongoing study, aiming to explore ways of improving productivity in home care.

A. Aim

Prior to the investigation the subject organization was seeking to improve productivity through the implementation of mobile ICT technology. The objective was to 'save time' on a scarce resource, caregiver time, to be used for coping with an increasing demand for services. With the use of a mobile platform-based technology (henceforth 'solution'), certain office tasks (e.g., charting) were to be transferred to the field, speeding up and improving the efficiency of the caregivers' administrative work routines.

In order to avoid investing in a solution that would not increase throughput, while simultaneously increasing operating expenses, the authors sought to investigate the potential solution's effect on system constraints.

B. Home Care Operations

Home care (or domiciliary care) refers to regular healthcare or supportive services provided in a customer's own home by a visiting caregiver. Home health care, meaning skilled nursing, is sometimes distinguished from home care, meaning non-medical care. Here they are treated jointly under the term home care. The services range from medical (e.g., health care and hospice) to supportive (e.g., bathing) and social services (e.g., transportation). The purpose of home care is to enable people in need of assistance to continue living in their own homes by improving, maintaining, or reducing the natural stagnation of their health conditions and autonomy. Although age is not a basis for service discrimination, the vast majority of the clientele is typically made up of older adults, whose autonomy is reduced.

The frequency of home visits varies with customers' specific needs, ranging from occasional monthly visits up to 4 visits per day, averaging 1.56 daily visits on weekdays in the organization that was studied. Although the absolute majority of the customer encounters constitute home care visits, occasionally services are provided over the phone, or a customer visits the office.

The demand for services is based on an evaluation, performed by the caregivers, typically a nurse and a social worker, in collaboration with the potential customers. During the process, a care plan is made out for the accepted customers. The care plan is revised biannually or when changes in a customer's condition so require.

The output is the volume of procedures or service units performed (the provider perspective). An outcome is a change in a patient's medical condition, which carries some health value (the patient perspective). The outcome (or value) is therefore the ultimate goal. The outcome cannot, however, be used as an operational indicator in situations where the outcome is significantly affected by factors beyond the service provider's control, such as patients' health behavior, placebo effects, or random events. The throughput is therefore the provider's contribution to the creation of outcomes (health value). It was decided that the most appropriate definition of throughput was not the number of customers served, but the service time produced. Defining throughput as the number of customers or visits may risk constructing incentives to maximize the volume of encounters, rather than the effectiveness of the service.

The schedule of the caregivers is based on the demand, as stated in the care plan. In home care, the demand for services is typically greatest in the morning [28]. The caregiver capacity is staggered into two 7.5-hour-long shifts. The morning shift workers start around 7-8 a.m., finishing around 3-4 p.m., while the evening shift staff comes on after 2 p.m., finishing as late as 10 p.m. Although there is a slight variation in the starting and finishing times, the caregiver capacity is quite stable throughout the shifts, with a clear drop between the morning and evening shifts, the capacity of the evening shift being approximately 25% of that of the morning shift.

C. Methods

The inquiry adopted both qualitative and quantitative methods. The qualitative part was conducted to gain a rich understanding of the service operations of the home care organization that was studied. It included interviews and regular meetings with management, an ethnographic study of the service process and work routines, and several workshops with staff. Some results from the qualitative part of the study were first presented in Groop et al. [1].

The management team that was interviewed included the head of Home Care, a service manager, two ICT specialists, and one financial specialist. The objective was to get an overview of the service production system and its dynamics. Once the quantitative study was in progress, regular meetings were held with the management team for the purpose of feedback and analysis of findings. For the ethnographic study, the first author observed a staff member at work for an entire shift. A total of three staff members were observed on three separate occasions. The first two were registered nurses working in separate teams, while the third was a foreman. The first workshop was held at an early stage of the investigation. Its objective was to strengthen the comprehension of everyday services through a facilitated discussion with representatives from all levels of the home care organization. Once the quantitative study was finished, four workshops were held to disseminate and validate the findings. The first two workshops included all the foremen,

while the latter two targeted two separate home care teams, chosen by the management.

The purpose of the quantitative part of the study was to quantify the home care operations, focusing on throughput and caregivers' workload. The investigation consisted of a longitudinal analysis of operational data. Data illustrating the distribution of throughput, i.e., the amount, duration, and timing of realized home care visits, were collected for a period of six weeks, from February 9th to March 22nd 2009. This enabled analysis of the caregivers' workload, i.e., services performed in terms of time, over a period of time.

The data consisted of two consecutive three-week scheduling periods, and they were chosen in collaboration with management. According to the management team, the chosen period exemplified the most "normal" conditions, as it was least affected by staffing exceptions as a result of holidays. The data represented a total of 43,716 home care encounters amounting to a total of 21,934 hours of service.

D. Participants

The organization that was studied, Espoo Home Care (EHC), is a large public health organization. It is responsible for providing statutory home care services for the City of Espoo, one of the largest municipalities in Finland. The organization is divided into service homes and field-based services (Regional Home Care). Since the potential technology implementation targeted only the field-based services [29], the service homes were excluded from the analysis. Some services, such as night-time care (10 p.m. to 7 a.m.), customer transportation, and meal and grocery deliveries, as well as care for certain severely disabled customers, are outsourced. These services were also excluded from the analysis. Services performed by temporarily leased caregivers, covering for absent employees, were included.

Regional Home Care employed a field-based staff of 326 (as of Feb 27th 2009), of whom 53 (16.2%) were part-time employees, and 19 primarily office-based foremen, who were trained registered nurses or social workers. Out of the field-based personnel, roughly 23% were registered nurses, 61% practical nurses, and 15% home care assistants. The home services provided during the six-week period included services rendered to 2587 customers, by 293 caregivers, for an average of 801 customers per day (STD 40) on weekdays, and 389 (STD 12) on weekends. This corresponds to 1189 (STD 55) customer encounters on weekdays and 671 (STD 44) on weekends.

The clientele of ECH consisted of both temporary customers, who exit the system once their health/autonomy improves, and regular customers, exiting the system either through death or transfer to a more comprehensive form of care, e.g., sheltered accommodation or long-term care. During the period of the study temporary customers received only 0.34% of the total services.

V. FINDINGS

This section first explains the process changes that certain technologies make possible, and provides reasons why these changes might help to improve productivity. After this, the TOC framework is adopted to evaluate whether the suggested process changes would in fact have the intended effect on productivity.

A. Time and Location Constraints

The home care service delivery process is field-based, which implies that there are constraints related to time and location [30], which have a great impact on the production flow. In the ECH the caregivers visit the office each morning, to collect their work list specifying which customers to visit, when, and which types of services to perform. The caregivers also pick up the keys to the homes of several of their customers, and company cars, if assigned one. Once the caregivers have completed their rounds, they return to the office to bring back the customers' house keys and to perform office tasks, such as data entry into the EMR (electronic medical record) (i.e., charting: visit summaries, updating patient information and care plans, placing customers' grocery and meal orders etc.). All of these activities are subject to time and location constraints that can be broken using technology. The time constraint means that certain tasks have to be performed in a certain order (e.g., getting the keys before entering a customer's home), and at certain times when the customer and caregiver meet. The location constraint means that certain services need to be performed at locations where the customers (home) or equipment (office computers) are located. The time and location constraints force caregivers to spend time moving between locations, rather than spending time serving their customers, which reduces throughput.

Mobile solutions that allow data entry and the retrieval of patient information to and from the electronic medical record (EMR) can reduce the need to visit the office. Data can be charted in the field rather than using a computer at the office, thereby eliminating a step in the process. This would be true if data entry were the only reason why caregivers visited the office. There may, however, be other time and location constraints. In this case, the caregivers also have to collect and return the customers' home keys. This causes another time and location constraint that offsets the potential of the mobile data entry and retrieval system. There are, however, technologies that allow a mobile platform to turn into a door-opening device, combining two technologies that together may have the power to break the constraint.

The constraints matrix (Figure 2) is an instrument that helps visualize the relationship between technologies and constraints, once these are identified. Figure 2 shows an example of the matrix based on the observations from the case study, including the previously presented time and location constraints that constitute major everyday obstacles

to a swift even flow [20]. There are other reasons for visiting the office as well, such as team meetings or collecting and sorting customers' medication. These activities, however, differ from the constraints used in the example in that they do not have to be performed every day, and can be scheduled.

The technologies are wireless mobile platform-based solutions. The technologies have different capabilities, each breaking some of the constraints. Technology 1 affects constraints 1, 2, and 4, while technology 3 only affects constraint 1. The technologies considered for implementation included software applications which make possible either one-way or two-way charting of the EMR, i.e., wireless data entry, or both wireless data entry and retrieval, as well as flexible mobile scheduling, i.e., mobile work lists. One technology was a Bluetooth-based solution for wireless door-opening. The solution can be integrated into any Bluetooth-equipped mobile device. The capabilities of the technologies are:

- Technology 1: two-way charting and scheduling
- Technology 2: two-way charting and scheduling
- Technology 3: one-way charting (data entry)
- Technology 4: Bluetooth-based wireless door-opening

Not all constraints have the same impact on the system. Some are more severe than others. It is likely that the home key constraint will have a lesser impact on the system's throughput than the mobile data entry and retrieval technology. The caregivers' information processing activities consume considerably more time than customers' home keys logistics. However, due to synergy advantage, together the technologies are bound to contribute more than the sum of their parts, as they can remove an entire process step.

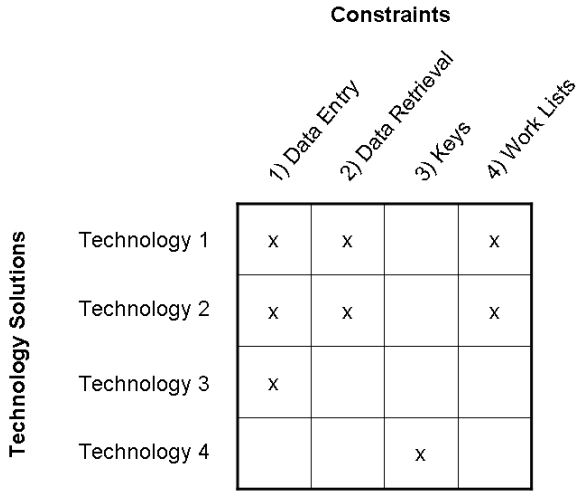


Figure 2. Constraints matrix showing which constraints the alternative technologies affect.

Unfortunately, not all technologies are equally easy to implement. The technology, and the ease or difficulty of implementation, is termed feasibility. It incorporates the following variables: a technology's usability; the cost of the investment; the process changes involved, and their ease of implementation. There may be considerable differences both in terms of operational feasibility and in the amount of education and training required. Technologies by themselves rarely have the capability to increase throughput and improve productivity. Their power lies in their ability to allow the redesigning of the way in which tasks are performed, so that more can be done with less.

For example, mobile data entry and retrieval technologies can allow home care caregivers to spend more value-adding time with their customers, improve the quality of the EMR information as a result of timely data entry, and remove the need for double data entry. Most activities will still need to be performed, only now at a more desirable time and place, which in turn may improve the operational flow and process throughput (quality improvement), or the same amount of customers can be treated with a smaller workforce (increased productivity).

When evaluating technologies it is imperative that the process redesign be accounted for. Thus, before evaluating new technologies on the basis of financial measures such as ROI and payback time, the ease of implementing the accompanying process changes should be evaluated. Even if the time and location constraints which currently force the caregivers to visit the office were to be removed, will the caregivers stop going there? There might be other reasons for visiting the office, such as social purposes. By implementing certain technologies the constraints can be eliminated, but the true benefits will not be realized before the operating procedures and practices change.

Goldratt et al. [31] explain that constraints force organizations to create rules (policies or routines) to cope with or work around exisiting constraints. Apparently, it is not uncommon for these rules to remain after the constraint has been removed. Failure to revise the rules has been shown to keep organizations from realizing the benefits of new technologies. The authors believe routine changes can be brought about by implementing suitable incentives.

B. Constraints vs. Non-Constraints: Timing Matters

The quantitative analysis of operational data showed the existence of considerable peaks in the workload. Figure 3 shows the fluctuation in the workload throughout the day, in terms of the total amount of service time. Roughly fifty percent of the services are performed during peak hours, from 8-11 a.m. This causes a peak time resource constraint during the morning rush hour, throughout which the system has trouble coping with demand.

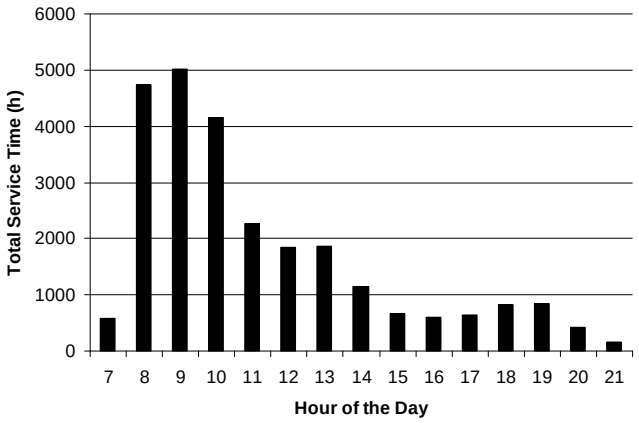


Figure 3. The caregivers' workload distribution. The figure shows the accumulated service time per hour of the day over a six-week period.

The intended mobile technology implementation sought to improve productivity by saving time on office tasks. However, the office tasks are performed in the afternoon during a period of low demand. Because of the uneven workload distribution, the time would be saved during hours of excess capacity (i.e., a non-constraint). TOC tells us that throughput can only be increased by alleviating or breaking the constraint, not by increasing the efficiency of non-constraints. In practice, saving time in the afternoon would not allow the organization to serve more customers. Consequently, implementing the solution in the current system is unlikely to save time during the peak time constraint.

By examining the home care operations through the TOC lens, we concluded that the sought technology implementation would not have the desired affect on productivity. The analysis also gave us the insight that the workload distribution was constraining the productivity of the system. However, once the workload is more level, alleviating the time and location constraints by implementing the technology is disposed to enable the organization to service more customers, improving productivity.

VI. DISCUSSION

Assessing the technologies from a traditional cost perspective – comparing the costs of the investment with its potential time savings multiplied by cost per hour – would not have provided a true representation of the effects on the system. First, reducing the required caregiver time in the afternoon does not reduce the organization's operating expenses, since labor is a fixed cost. The capacity remains unaffected, but capacity utilization is reduced, which spreads the fixed costs over less service time. Second, an investment in technology increases operating expenses. Third, in the current mode of operation, as *time saved in the afternoon cannot be used to treat more customers* (there is already overcapacity in the afternoon), no additional throughput is gained. Consequently, there is no added throughput to offset the increase in operating expenses. This reduces productivity, the complete opposite of what the planned investment sought to achieve.

A. Limitations of the Study

As a result of a lack of accurate data on the timing and duration of office tasks – unlike customer encounters, back-office activities, such as office tasks and transit, are not recorded – the contribution of office activities to the total workload could not be accurately quantified.

The capacity utilization rate (CUR) of the caregivers would have been a more appropriate measure than throughput for analyzing the workload. The distribution of throughput (i.e., service time) was therefore used as an approximation of the workload. From a previous activity-based costing investigation it was known that the office tasks consume less than 20 percent of the caregivers' total working time. Transit accounts for approximately 12 percent, and is naturally distributed alongside the throughput. At the time of the study, the ratio of service time to back-office time was less than forty percent. From this it can be construed that currently there is excess capacity in the afternoon, even though it is not quite as remarkable as Figure 3 would suggest, as the figure does not account for back-office activities.

Because of the problem-solving nature of the investigation, only one organization was studied. Both the literature [28] and the authors' experience, however, suggest that the morning demand peak is a common but poorly understood characteristic of home care operations.

B. A Note on Financial Analysis

When constraints and their priorities are understood, organizations can proceed to financial evaluation. Financial evaluation methodology has been widely discussed in the literature [32][33][34][35][36] and basic concepts can be applied to healthcare technology evaluations. We briefly present two commonly used methods, cost-benefit analysis and cost-effectiveness analysis, as examples of basic financial evaluation tools. Cost-benefit analysis (CBA) uses a monetary frame of reference to evaluate both outcomes and costs. Both the costs of interventions and the values of outcomes are assessed in terms of money. This analysis is particularly useful if the outcomes exceed costs and the solution with the largest net benefit (outcomes subtracted for costs) should be selected. Unlike CBA, the focus of cost-effectiveness analysis (CEA) is on the non-monetary outcomes of an intervention, such as a health improvement. CEA compares the cost of alternative (intervention) outcomes. "Alternatives are calculated and presented in a ratio of incremental cost to incremental effect" [37]. CEA is therefore more suitable for assessing technologies that

aspire to improve health outcomes than for evaluating process technologies.

Depending on the definition of throughput (money, visits, customers, service time etc.), these methods can be used to evaluate the financial implications of a change in throughput. The focus should, however, be on the throughput of the primary constraint, not the non-constraints'. Following TA, technologies could be assessed on the basis of their ability to increase the measure 'throughput per constraint minute' [3].

C. When is the TOC Approach Appropriate?

Prioritizing throughput over cost is particularly suitable for organizations that have sufficient demand to absorb the increased throughput. If, however, demand is a constraint (market constraint [12]), organizations should shift the focus to producing the same services with fewer resources, i.e., maintaining throughput while reducing operational expenses. Subject to a market constraint, the bottleneck resource that governs throughput, to which the rest of the system should be subordinated, is the resource with the highest utilization after market demand is satisfied [38]. As such, the TOC approach can help improve productivity even in the absence of an actual resource constraint.

The framework provides a rough, easy, and simple approach to ranking alternative technologies, according to their ability to affect the true operational limitations, i.e., the constraints. It is advisable to use this framework as a first step in the right direction. The suggested approach stresses that increasing throughput will reduce costs, while reducing costs will ultimately reduce throughput. Therefore the primary focus of technology assessment should be on the ability of technologies to increase throughput rather than reduce costs, so that existing resources can be used to their full potential.

VII. Conclusion and Future Work

New technologies are traditionally assessed on the basis of simplified financial estimates, with the operational impact being overlooked. This paper suggests an alternative approach based on Operations Management (OM), following the theory of constraints (TOC). The TOC approach is more pragmatic and suitable for technology evaluation in a fast-paced and growing health technology market. It holds that the performance of the system can only be increased by improving the performance of the primary constraint. Therefore technologies should be assessed on the basis of their ability to improve the performance of the constraint. The suggested approach further stresses that prioritizing throughput over cost will reduce costs, while focusing on cost reduction will ultimately reduce throughput. Once the operational aspects have been assessed, financial methods are bound to be more reliable and useful.

The empirical study showed that the time and location constraints were not the organization's primary constraint, as originally conceived. Alleviating or removing these constraints would therefore not have the desired affect. This provided the insight not to invest in the planned mobile solution, under the current mode of operation. The analysis further prompted the need to alleviate the impact of the peak time workload in Espoo Home Care by leveling service provision. An investigation of the causes behind the uneven workload distribution is currently taking place.

Suggested future work includes further development of the methodology and empirical testing of the presented framework in other settings, particularly those in which a technology is actually implemented. Comparing the use of this instrument to evaluations performed with other methods would be a valuable contribution.

Further testing and reporting on the suitability of the suggested framework for evaluating investment decisions outside healthcare is encouraged. Although TOC aspires to be the basis for all decision making, there seems to be a lack of published empirical studies on using it for this type of investment analysis.

Author Contributions

J.G., K.R, and P.L. were responsible for the conception of the study and drafted the manuscript. J.G. performed the empirical investigation, gathering and analyzing the data.

Acknowledgments

The authors wish to express their gratitude to the staff of Espoo Home Care, who were instrumental in making the study possible. In addition, Teemu Mustonen of Jyväskylä University and Ecomond Ltd. deserves special recognition for assisting in processing the quantitative raw data. The authors would further like to thank Professor Erkki Vauramo and Dr. Anu Helkkula of Aalto University, as well as the anonymous reviewers, for their valuable comments, which greatly helped to improve this paper.

During the preparation of the manuscript, J.G. and K.R. were visiting researchers at Stanford University (California). J.G. was affiliated with the Mechanical Engineering Design Group and K. R. was affiliated with the Stanford Prevention Research Center (SPRC). Both departments deserve acknowledgment for providing a stimulating research enviroment.

This study (PARETO) was funded by the European Regional Development Fund (ERDF).

References

[1] P. J. Groop, K. H. Reijonsaari, and P. M. Lillrank, "A Theory of Constraints Approach to Health Technology Assessment," in *eTelemed 2010, Second International Conference on eHealth, Telemedicine, and Social Medicine*, St. Maarten, Netherlands Antilles, pp. 147-152, February 10-16, 2010.

[2] J. Concato, N. Shah, and R. I. Horwitz, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *The New England Journal of Medicine*, vol. 342, no. 25, pp. 1887-1892, Jun. 2000.

[3] E. Goldratt, *What is this thing called theory of constraints and how should it be implemented?* Croton-on-Hudson, NY: North River Press, 1990.

[4] M. Gupta and L. Boyd, "Theory of constraints: A theory for operations management," *International Journal of Operations and Production Management*, vol. 28, no. 10, pp. 991-1012, 2008.

[5] A. M. Breen, T. Burton-Houle, and D. C. Aron, "Applying the Theory of Constraints in Health Care: Part 1-The Philosophy.," *Quality Management in Health Care*, vol. 10, no. 3, p. 40, 2002.

[6] M. G. Hunink, "In search of tools to aid logical thinking and communicating about medical decision making," *Medical Decision Making*, vol. 21, no. 4, p. 267, 2001.

[7] R. A. McNutt and M. C. Odwazny, "The theory of constraints and medical error: a conversation with Robert A. McNutt," *Quality Management in Healthcare*, vol. 13, no. 3, p. 183, 2004.

[8] J. Schaefers, J. Colin, R. Aggoune, and M. Kucina, "A contribution to performance measurement in the healthcare industry: the industrial point of view," *International Journal of Business Performance Management*, vol. 9, no. 2, pp. 226–239, 2007.

[9] S. Sadat, Theory of constraints for publicly funded health systems. Dissertation: University of Toronto (Canada), 2009.

[10] M. Gupta and J. Kline, "Managing a community mental health agency: a theory of constraints based framework," *Total Quality Management & Business Excellence*, vol. 19, no. 3, pp. 281–294, 2008.

[11] J. Motwani, D. Klein, and R. Harowitz, "The Theory of Constraints in Services: Part 2 - Examples from Health Care," *Managing Service Quality*, vol. 6, no. 2, pp. 30 - 34, 1996.

[12] B. Ronen, J. S. Pliskin, and S. Pass, *Focused Operations Management for Health Services Organizations*. San Francisco: Jossey-Bass, 2006.

[13] J. Wright and R. King, *We All Fall Down: Goldratt's Theory of Constraints for Healthcare Systems*. Great Barrington, MA: North River Press, 2006.

[14] J. E. Wennberg, *Tracking Medicine: A Researcher's Quest to Understand Health Care*. New York: Oxford University Press, 2010.

[15] E. Goldratt and J. Cox, *The Goal*. Croton-on-Hudson, NY: North River Press, 1984.

[16] S. U. Rahman, "Theory of constraints: a review of the philosophy and its applications," *International Journal of Operations and Production Management*, vol. 18, pp. 336-355, 1998.

[17] J. Motwani, D. Klein, and R. Harowitz, "The Theory of Constraints in Services: Part 1 - The Basics," *Managing Service Quality*, vol. 6, pp. 53–56, 1996.

[18] J. Kujala, P. Lillrank, V. Kronström, and A. Peltokorpi, "Time-Based Management of Patient Processes," *Journal of Health Organization and Management*, vol. 20, no. 6, pp. 512–524, 2006.

[19] J. H. Blackstone, "Theory of Constraints - A Status Report.," *International Journal of Production Research*, vol. 39, no. 6, pp. 1053-1080, Apr. 2001.

[20] R. Schmenner and M. Swink, "On theory in operations management," *Journal of Operations Management*, vol. 17, no. 1, pp. 97-113, 1998.

[21] R. Schmenner, "Manufacturing, service, and their integration: Some history and theory," *International Journal of Operations and Production Management*, vol. 29, no. 5, pp. 431-443, 2009.

[22] D. Jacob, S. Bergland, and J. Cox, *Velocity: Combining Lean, Six Sigma and the Theory of Constraints to Achieve Breakthrough Performance-A Business Novel*. New York: Free Press, 2009.

[23] A. Morton and J. Cornwell, "What's the difference between a hospital and a bottling factory?," *British Medical Journal*, vol. 339, no. 20, p. b 2727, 2009.

[24] E. Goldratt and R. E. Fox, *The Race*, 1st ed. Croton-on-Hudson, NY: North River Press, 1986.

[25] B. Ronen and Y. Spector, "Managing system constraints: a cost/utilization approach," *International Journal of Production Research*, vol. 30, no. 9, pp. 2045–2061, 1992.

[26] L. Boyd and M. Gupta, "Constraints management: What is the theory?," *International Journal of Operations & Production Management*, vol. 24, no. 4, pp. 350 - 371, 2004.

[27] E. Goldratt, *The haystack syndrome: Sifting information out of the data ocean*. North River Press Great Barrington, MA, 1990.

[28] P. Eveborn, P. Flisberg, and M. Rönnqvist, "Home Care Operations," *OR/MS Today*, vol. 31, no. 2, pp. 38–43, 2004.

[29] S. Agnihothri, N. Sivasubramaniam, and D. Simmons, "Leveraging technology to improve field service," *International Journal of Service Industry Management*, vol. 13, no. 1, pp. 47–68, 2002.

[30] K. H. Ilvonen, P. J. Groop, and P. M. Lillrank, "Mobile Services Provide Value by Decoupling the Time and Location Constraints in Healthcare Delivery," in *eTelemed 2009, First International Conference on eHealth, Telemedicine, and Social Medicine*, Cancun, Mexico, pp. 216-219, February 1-7, 2009.

[31] E. Goldratt, E. Schragenheim, and C. A. Ptak, *Necessary But Not Sufficient*. Croton-on-Hudson, NY: North River Press, 2000.

[32] T. T. Edejer et al., "WHO guide to cost-effectiveness analysis," *WHO, Geneva*, 2003.

[33] M. Johannesson and B. Jönsson, "Economic evaluation in health care: is there a role for cost-benefit analysis?" *Health Policy*, vol. 17, no. 1, p. 1, 1991.

[34] H. E. Klarman, "Application of cost-benefit analysis to health systems technology," *Journal of Occupational and Environmental Medicine*, vol. 16, no. 3, p. 172, 1974.

[35] K. E. Warner and R. C. Hutton, "Cost-benefit and cost-effectiveness analysis in health care: Growth and composition of the literature," *Medical Care*, vol. 18, no. 11, pp. 1069–1084, 1980.

[36] A. E. Boardman, D. H. Greenberg, A. R. Vining, and D. L. Weimer, Cost-benefit analysis: concepts and practice, 3rd Ed. Upper Saddle River, NJ: Prentice Hall 2006.

[37] J. Brazier, J. Ratcliffe, and A. Tsuchiya, *Measuring and valuing health benefits for economic evaluation*. New York: Oxford University Press, 2007.

[38] J. Balakrishnan, C. H. Cheng, and D. Trietsch, "The Theory of Constraints in Academia: Its Evolution, Influence, Controversies, and Lessons," *Operations Management Education Review*, vol. 2, pp. 97–114, 2008.

Nonlinear Law Spectral Technique to Analyze White Spot Syndrome Virus Infection

Mario Alonso Bueno-Ibarra
Departamento de Biotecnología Agrícola,
Centro Interdisciplinario de Investigación para el Desarrollo Integral Regional (CIIDIR - Sinaloa).
Blvd. Juan de Dios Bátiz Paredes #250,
Col. San Juachín,
Guasave, Sinaloa C.P. 81101.
mbueno@ipn.mx

María Cristina Chávez-Sánchez
Laboratorio de Histología,
Centro de Investigación en Alimentación y Desarrollo A.C. (CIAD - Mazatlán).
Av. Sábalo-Cerritos S/N, Estero del Yugo,
Mazatlán, Sinaloa, México, C.P. 82000.
marcris@ciad.mx

Josué Álvarez-Borrego
División de Física Aplicada, Departamento de Óptica,
Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE).
Carretera Ensenada-Tijuana No. 3918,
Fraccionamiento Zona Playitas.
Ensenada, Baja California, México, C.P. 22860.
josue@cicese.mx

Abstract **- In this paper a novel spectral technique based on K-Law Fourier nonlinear methodology is developed to classify White Spot Syndrome Virus inclusion bodies found in images from infected shrimp tissue samples. Shrimp culture is expanding in the world due to their high demand and price. This rapid increase in cultured shrimp production was achieved by geographical expansion and technological advances in reproduction in captivity of the white shrimp *Penaeus* vannamei, larval rearing, artificial diet and intensification in growth out systems. However, diseases are one of the major constraints for the sustainable increase of shrimp production. The white spot syndrome virus is a pandemic disease where frequently 100% mortality may occur within 2-3 days. However, several techniques have been implemented and developed for viral and bacterial analysis and diagnostic's tasks; histology is still considered the common tool in medical and veterinary fields. The slide images were acquired by a computational image capture system and a new spectral technique by the development of a spectral index is done to obtain a quantitative measurement of the complexity pattern found in White Spot Syndrome Virus inclusion bodies. After analysis the results show that inclusion bodies are well defined in a clear numerical fringe, obtained by the calculation of this spectral signature index.**

Keywords - WSSV; image processing techniques; inclusion bodies; shrimp; virus

I. INTRODUCTION

Several techniques and methods including microscopic observation under light, dark field, phase contrast microscope, bioassay, transmission electron microscopy, immunological, molecular and histopathological methods have been developed for viral and bacterial penaeid shrimps diagnostics [1][2]; these can be divided in traditional morphological pathology, bioassay, microbiology and serology and molecular methods such as PCR, however histology is still considered the common tool in medical and veterinary for research and diagnostics tasks [3][4][5]. The contribution of aquaculture to world supply of fish, crustaceans, mollusks and other aquatic animals has continued to grow, and has gone from a 3.9% of total production in weight in 1970 to 36.0% in 2006. The world's supply of crustaceans by aquaculture has grown rapidly in the last decade, and has reached 42% of the world production in 2006 and, in that same year, provided the 70% of shrimps and prawns (peneidos) produced worldwide [6]. *Penaeus vannamei* is the most important shrimp species in terms of aquaculture production and is naturally present along the Pacific coast of Central and South America [7].

This species was originally cultured in North, Central and South American countries but at the end of the 1970s, this species was introduced in Asia [8]. In spite of this important increase of shrimp production, the shrimp culture industry has been affected since the 80´s with different important diseases [4], being the viruses the most significant pathogens in shrimp. Viral diseases have caused considerable losses of production and jobs, reduced earning, export restrictions, failure and closing of business and decreased confidence of consumers [9].

The white spot syndrome virus (WSSV) is considered as a serious pathogen and is actually a pandemic disease causing important effects in production in many countries with the consequent social impact of many Asian and American countries. México traditionally is considered a privilege country where extensive commercial shrimp aquaculture can be exploited, due its weather and location close to the biggest consumer of the world, United States of America.

Since 2000, in México, shrimp producers from the Northwest of the country have been affected by WSSV. Sinaloa and Nayarit states for example, were reflected their

losses by the reduction of exportations from 30.1 million USD in 2000 compared to the 44.8 million USD in 1999, the losses amount were approximately 14.7 million USD just in one production year, thus producers after this year have been taking actions to control the WSSV disease to reduce their impact [10].

Today the virus continues affecting the production causing the closure of many businesses and affecting the capital of many investors. It is reported by example that in the first half of 2009, the virus has caused mortalities of 70-80 per cent in affected farms in Sinaloa and in Sonora the virus was dispersed in 2008 to areas of the state that had remained free of the pathogen (Personal communication with the Aquatic Health Committees of Sinaloa and Sonora).

WSSV infections have been detected in various tissues and organs, hemolymph, gills, stomach and body cuticular epithelium, hematopoietic tissues, lymphoid organ, antennal glands, connective tissues, muscle tissues, hepatopancreas, heart, midgut, hindgut, nervous tissues, compound eyes, eye, testes and ovaries of naturally and experimentally infected shrimp [11]. WSSV can spread and infect shrimps of any stage of grow-out, asymptomatically affecting all life cycle stages, from eggs to broodstock. Once the clinical signs are developed, mortality can reach 100% in 3 days. The causative agent of the disease is an ovoid or ellipsoid to bacilliform in shape double-stranded DNA virus (120-150 nm in diameter and 270-290 nm in length), which genus is Whispovirus, within the family Nimaviridae [12][13] .

There is no efficient approach to control this disease, thus, the need for rapid, sensitive diagnostic methods led to develop new alternative techniques in different fields of knowledge. The selection of a method is dependent on the purpose, e.g., histopathology is used to determine the affected organs, the levels of affectation, the pathological changes in cells, tissues and organs and computing optic disciplines can be of support to these conventional methods. Histology makes possible to analyze pathological changes in several tissue cells and allow the pathogen identification, which are sometimes difficult to recognize with other alternative techniques.

For this kind of analysis the method involves several steps to obtain the final sample, which contains a shrimp tissue slice where the inspection is conduced; this tissue slice has a thickness of 1-5 µm, stained with hematoxilin - eosin necessarily to make the examination under microscope. WSSV infection is commonly seen in cuticular epithelial cells and connective tissue cells of the stomach, carapace and gills. However it is also seen in antennal gland, lymphoid organ, hematopoietic tissue and phagocytes of the heart, some WSSV samples are shown in Figure 1.

Infected cells typically have hypertrophied (enlarged) nuclei containing a single intranuclear inclusion. Inclusions at the beginning are eosinophilic and sometimes are separated by a clear halo beneath the nuclear membrane; these are known as Cowdry type A inclusions. Later inclusions become lightly to deeply basophilic and fill the entire nucleus [14], as shown in Figure 2.

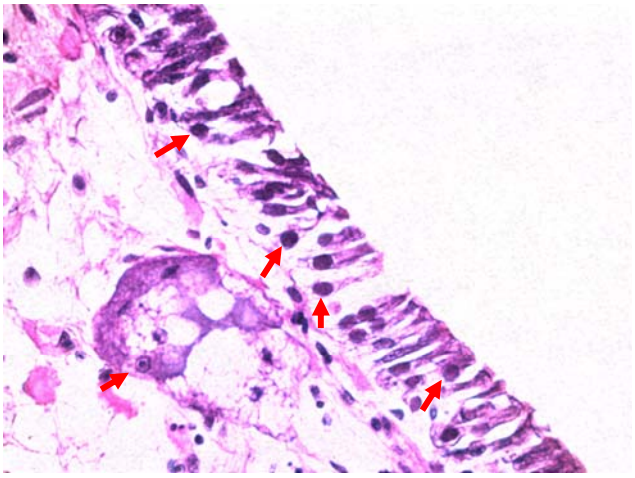


Figure 1. Arrows show samples of connective tissue cells presenting the WSSV infection.

There is no efficient approach to control this disease, however the WSSV early detection can prevent the shrimp's mortality by application of different strategies, one of them may be the RNA genetic technique and it can be constituted as a new therapeutic strategy to control the WSSV infection [15]. Thus, the need for rapid, sensitive diagnostic methods led to develop new alternative techniques in different fields of knowledge like computing optic disciplines that can be of support to conventional methods.

Several optic and computational techniques were developed to recognize these kinds of biological patterns, the analysis of inclusion bodies is determinant of the virus presence, e.g., the color correlation approach used to analyze and recognize the presence of IHHNV inclusion bodies by histological samples from 35 mm transparencies digitalized with a flatbed scanner [16].

The aim of this paper is to extent the development of a new technique to classify the WSSV basophilic and Cowdry type A inclusion bodies, acquired from histological digitalized images from infected shrimps samples by the analysis applied over WSSV sample's slides, based in the application of Fourier spectral filtering techniques, such as K-Law nonlinear filter technique [1].

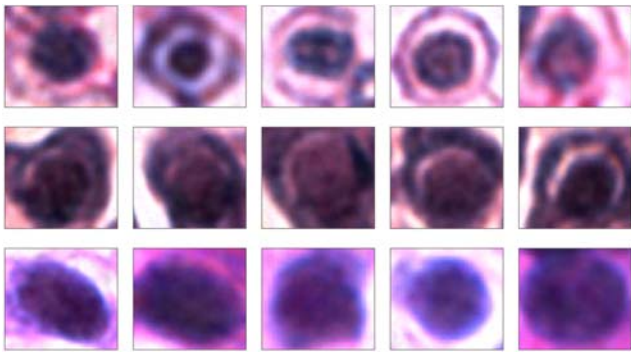


Figure 2. Typical infected shrimp's cells with WSSV basophilic and Cowdry A type inclusion bodies.

These Fourier spectral and color correlation techniques have been demonstrated the capability of analyze important characteristics from viruses and pathogens [16][17][18], including applications in several fields [19][20][21][22].

Therefore, section II describes the core basis of the methodology and the equipment used to obtain the images from infected shrimp tissues; section III presents the results obtained with the spectral signature index developed; additionally the statistical analytic values obtained from this index and the experiments carried out from the digitalized images previously acquired are included; finally in section IV the possible future work that needs to be done to enrich this research is discussed.

II. MATHERIALS AND METHODS

Development of a new technique to analyze and classify WSSV inclusion bodies is divided in five subsections; subsection A describes how the shrimp samples were prepared; subsection B describes the equipment used for image capturing; subsection C describes how had been determined the best spatial color channel function from the multispectral images where the WSSV texture measurement and tissue analysis are carry out by this technique; subsection D explains the mathematical basis used for this technique; finally subsection E describes the steps involved in the obtaining of the classification by the signature index proposed.

A. Virus sample preparation

Experimental shrimps were obtained from a farm located in the state of Sinaloa, México; transported live to the laboratory to be fixed in Davidson's solution; after 24 h, the fixative was discarded and shrimps were preserved in 50% alcohol solution until they were ready to be processed by conventional histology techniques [2][3][5][13] to obtain the final shrimp tissue histological sample slides, as shown in Figure 3, afterwards they were ready to be examined under microscope.

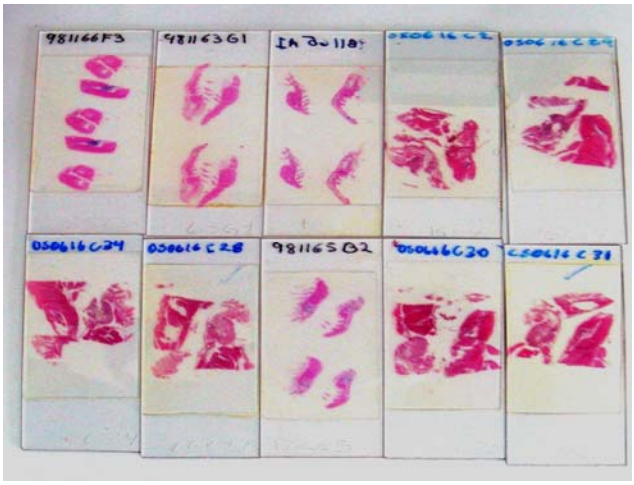


Figure 3. Shrimp tissue histological sample slides ready to be analyzed under microscope.

Several types of WSSV inclusion bodies were selected from cuticular epithelium, connective tissue and abdomen tissue and were digitalized to obtain a filter bank.

Figure 4. Leica microscope model DMRXA2 equipped with a RGB color 3.2 mega pixel digital camera (Leica model DC300) attached to a 2.5 GHz PC Pentium IV.

B. Digitalized images capture

The WSSV sample slide images were acquired by a novel computational image capture prototype system to enhance the digitalized images with novel autofocus and fusion techniques developed [23][24], running inside a 2.5 GHz PC Pentium 4 with 1 GByte RAM and 80 GBytes HD attached to Leica microscope (model DMRXA2) equipped with a RGB color digital camera (Leica model DC300), as shown in Figure 4.

A set of 168 microscope images were acquired from the shrimp's tissues by 60x objetive with a 2088 x 1550 pixels color resolution digital camera; each representative field can contains about an average of 30 to 60 approximately inclusion bodies depending of the level of WSSV infection. Afterwards, a set of 870 WSSV inclusion body images were selected to build a filter bank with 100 most representative WSSV images obtaining the intensity spatial domain matrix data of each WSSV image to be analyzed; thus multispectral function $f^{\lambda}(x,y)$ is defined for every pixel coordinates x and y on digitalized images, where $\lambda=\{\lambda_R,\lambda_G,\lambda_B\}$ acquired by a CCD's digital camera with range $[0, 255]$ and red (R), green (G) and blue (B) are channels in RGB color space representation.

C. Multispectral analysis for the best spatial function channel determination

Let us introduce some useful definitions and functions: $f_1^{\lambda}, f_2^{\lambda}, f_3^{\lambda},.., f_w^{\lambda}$ are a multispectral filter bank of W captured images of size $N\times P$ pixels from inclusion body samples taken; $f_w^{\lambda}(x,y)$ is the captured image matrix with pixels (x,y) in the w^{th} inside filter bank images, where $x=1,..,P$, $y=1,..,N$ and $w=1,..,W$.

Each inclusion body sample image $f_w^{\lambda}(x,y)$ digitalized can be decomposed by their corresponding RGB

$\lambda = \{\lambda_R, \lambda_G, \lambda_B\}$ channels; thus the three intensity matrix data are $f_w^{\lambda_R}(x,y)$, $f_w^{\lambda_G}(x,y)$ and $f_w^{\lambda_B}(x,y)$ respectively.

Let $\hat{\mathrm{H}}$ be a vector of real numbers $\mathbb{R}$, where $\hat{\mathrm{H}} \in \mathbb{R}$ with T elements, whose elements are sorted in ascending order with respect to their values, the maximum function $MAX(\hat{\mathrm{H}})$ and the integer function $\lfloor \eta \rfloor$ of a number can be expressed respectively, like

$$MAX(\hat{\mathrm{H}}) = \left\{ h_{\mathrm{T}} \mid h_i \le h_{i+1}, h_i \in \hat{\mathrm{H}}, i = 1,2,\ldots,(\mathrm{T}-1) \right\} \quad (1)$$

$$\lfloor \eta \rfloor = \left\{ \delta \mid \delta \in \mathbb{Z}, \eta \in \mathbb{R}, \delta \le \eta \le \delta + 1 \right\} \quad (2)$$

where $\mathbb{Z}$ represent the set of whole numbers.

Every intensity matrix channel of each inclusion body from the filter bank are analyzed by taking a intensity profile vector set $\left\{\xi_q^{\lambda}\right\}_w$ where $q = 1,..,Q$ vectors and $\left\{\xi_q^{\lambda}\right\}_w \in f_w^{\lambda}(x,y)$, thus $\left\{\xi_q^{\lambda}\right\}_w$ can be defined by

$$\left\{\xi_q^{\lambda}\right\}_w = f_w^{\lambda}(x, \zeta_{Vt}), \quad for\ Vt = -\left\lfloor \tfrac{Q}{2} \right\rfloor, .., \left\lfloor \tfrac{Q}{2} \right\rfloor \quad (3)$$

where $Vt \in \mathbb{Z}$, $\zeta_{Vt} = \left\lfloor \tfrac{N}{2} \right\rfloor + Vt$ and $x = 1,..,P$.

Let $\widehat{\Psi}^{\lambda}$ be a vector of mean values of intensity profile vector set, on each channel, where $\widehat{\Psi}^{\lambda} \in \mathbb{R}$, using (3) these mean values can be calculated as

$$\widehat{\Psi}_w^{\lambda} = \frac{\sum_q \left\{\xi_q^{\lambda}\right\}_w}{Q}. \quad (4)$$

Figure 5 shows the intensity profile vector set $\left\{\xi_q^{\lambda}\right\}_w$ graphics calculated from each WSSV channel using (3), afterwards, using (4) can be obtained the pattern measurement of every WSSV channel, $\widehat{\Psi}^{\lambda_R}$, $\widehat{\Psi}^{\lambda_G}$ and $\widehat{\Psi}^{\lambda_B}$ respectively. Calculating the maximum value MV_{Ψ} of the mean values vectors $\widehat{\Psi}^{\lambda}$ can be obtained the best spatial matrix data from where the WSSV inclusion body pattern is analyzed, MV_{Ψ} can be obtained by the following expression

$$MV_{\Psi} = MAX\left\{\widehat{\Psi}_w^{\lambda_R}, \widehat{\Psi}_w^{\lambda_G}, \widehat{\Psi}_w^{\lambda_B}\right\} \quad for \quad w = 1,..,W, \quad (5)$$

therefore the channel source from MV_{Ψ} has a maximum value indicates the best matrix data, as shown in Figure 6.

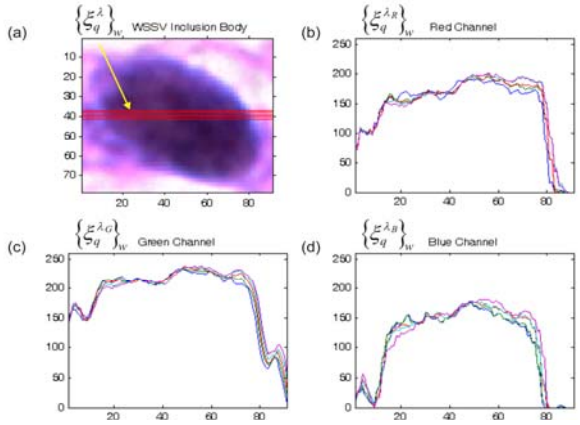


Figure 5. WSSV inclusion body analysis to get the best information channel, (a) intensity vector set, (b) red channel intensity profile, (c) green channel intensity profile and (d) blue channel intensity profile.

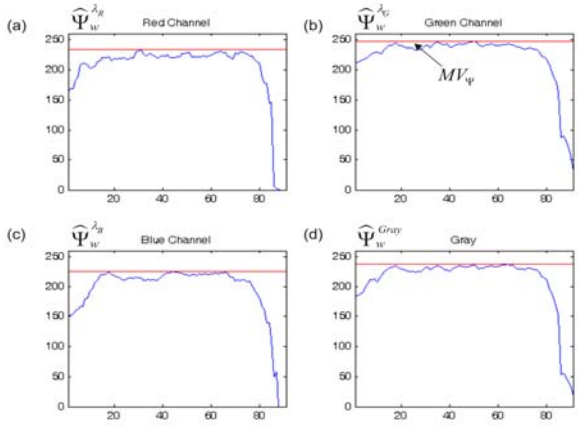


Figure 6. After analyzing the WSSV filter bank and obtaining the maximun value from intensity profile, (a) shows mean values from red channel, (b) shows green channel with the maximun value, (c) shows mean values from blue channel and (d) shows a RGB conversion to gray space with the mean values from its intensity profile.

D. WSSV spectral signature index classifier

The spectral signature *SSF* can be defined like a function to obtain the spectral properties of $f_w^{\lambda}(x,y)$ from a specific RGB channels $\{\lambda_R, \lambda_G, \lambda_B\}$; whereas the spectral signature index i^{ss} was developed to get a quantitative measurement of the inclusion bodies complexity pattern.

Let i^{ss} be defined like scalar number valued $i^{ss} \in \Re^{+}$ to measure the spectral frequency properties found in $f_w^{\lambda}(x,y)$ obtained by *SSF* function, let $I^{\lambda_G}(x,y)$ be the intensity matrix data obtained by $f_w^{\lambda}(x,y)$ green channel, where WSSV inclusion bodies characteristics are protruded.

Let function $I^{\lambda_G}_{Contour}(x,y) \in [0,1]$ be the WSSV inclusion body contour of the function $I^{\lambda_G}(x,y)$ calculated by the application of active contour technique (also known as a "snake") as follows

$$I^{\lambda_G}_{Contour}(x,y) = -\left|G_\sigma(x,y) * \nabla^2 I^{\lambda_G}(x,y)\right|^2, \quad (6)$$

where ∇^2 is the Laplacian operator, $G_\sigma(x,y)$ is a Gaussian standard deviation σ, afterwards the edge is obtained by the zero-crossings of $G_\sigma(x,y) * \nabla^2 I^{\lambda_G}(x,y)$ in the Marr-Hildreth theory [25].

Morphological reconstruction has a broad spectrum of several functions, e.g., like those to filling holes, defining the function $I^{\lambda_G}_{WSSV-Seg}(x,y)$ to be the image segmentation of WSSV inclusion body function $I^{\lambda_G}(x,y)$, it can be obtained by filling the area of WSSV inclusion body contour calculated by (6), let $B_{in} = \{(x,y)\}$ be the pixel coordinates set delimited by the area of the function $I^{\lambda_G}_{Contour}(x,y)$, then the pixel coordinates set can be obtained by $B_{in} = \left\{(x,y) \middle| \; inside\; I^{\lambda_G}_{Contour}(x,y) = 1\right\}$, afterwards the general segmentation function $I^{\lambda_G}_{Seg}(x,y)$ can be defined as follows

$$I^{\lambda_G}_{Seg}(x,y) = \begin{cases} 1 & \left\{\forall(x,y) \middle| \varphi(x,y) \in \gamma_{in}\right\} \\ 0 & otherwise \end{cases}, \quad (7)$$

where $I^{\lambda_G}_{WSSV-Seg}(x,y) = I^{\lambda_G}_{Seg}(x,y)$, $\varphi(x,y) = I^{\lambda_G}_{Contour}(x,y)$ and $\gamma_{in} = B_{in}$.

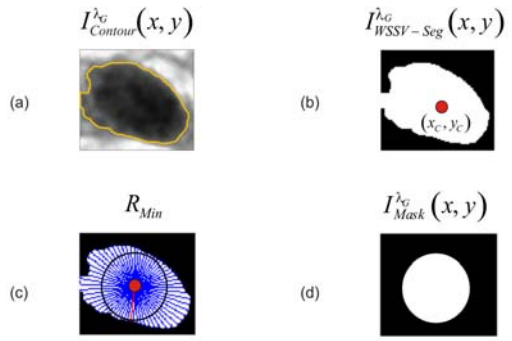


Figure 7. Segmentation steps, (a) shows the contour from WSSV inclusion body, (b) shows the centroid coordinates and image segmentation, (c) shows the calculation of minimum radius and (d) shows the circular mask to apply into its WSSV pattern analysis.

Once WSSV inclusion body segmented function is obtained, it is necessarily to make the analysis just inside of the WSSV inclusion body pattern, thus a binary circular function $I^{\lambda_G}_{Mask}(x,y)$ is calculated inside the segmented function $I^{\lambda_G}_{Seg}(x,y)$.

The center of the circular mask function is calculated by obtaining the centroid (x_C, y_C) of the WSSV inclusion body segmentation $I^{\lambda_G}_{WSSV-Seg}(x,y)$, the coordinates (x_C, y_C) can be calculated by the following pair equations as

$$x_C = \frac{\sum_{i=1}^{P}\sum_{j=1}^{N}(i) * I^{\lambda_G}_{WSSV-Seg}(i,j)}{\sum_{i=1}^{P}\sum_{j=1}^{N} I^{\lambda_G}_{WSSV-Seg}(i,j)} \quad (8)$$

and

$$y_C = \frac{\sum_{i=1}^{P}\sum_{j=1}^{N}(j) * I^{\lambda_G}_{WSSV-Seg}(i,j)}{\sum_{i=1}^{P}\sum_{j=1}^{N} I^{\lambda_G}_{WSSV-Seg}(i,j)}. \quad (9)$$

The minimum radius R_{Min} of the circular binary mask where the WSSV inclusion body pattern will be analyzed centered on (x_C, y_C), can be obtained defining the function $R(\theta)$ used to calculate the magnitude of a vector beginning in coordinates (x_C, y_C) until reach the edge of $I^{\lambda_G}_{WSSV-Seg}(x,y)$ in θ direction as follows

$$R(\theta) = \left\{ r \middle| \begin{array}{l} x^* = r \cdot cos(\theta);\; y^* = r \cdot sin(\theta) \\ r = r+1 : r \in \mathbb{Z}^+ \quad if\; I^{\lambda_G}_{WSSV-Seg}(x_c + x^*, y_c + y^*) = 1 \end{array} \right\}. \quad (10)$$

Afterwards doing $r_1 = R(\theta)$ and $r_2 = R(\theta + \Delta\theta)$ where a counterclockwise angle increment is $\Delta\theta$, thus the minimum radius R_{Min} can be defined by

$$R_{Min} = \begin{cases} r_1 & if\; r_1 \le r_2 \\ r_2 & otherwise \end{cases}, \quad (11)$$

then using (10) and (11) for $\theta = 0, ..., 2\pi - \Delta\theta$ is obtained R_{Min}. Let the function $Circ(R_{Min})$ be the circle created by the rotation of R_{Min}, then the $I^{\lambda_G}_{Mask}(x,y)$ circular binary mask function using (7) is obtained by $I^{\lambda_G}_{Mask}(x,y) = I^{\lambda_G}_{Seg}(x,y)$ doing $\varphi(x,y) = Circ(R_{Min})$ and $\gamma_{in} = \left\{(x,y) \middle| \; inside\; Circ(R_{Min}) = 1\right\}$, as shown in Figure 7.

Let the function $I_{IB}^{\lambda_G}(x,y)$ be the intensity matrix data resulted after the application of the $I_{Mask}^{\lambda_G}(x,y)$ circular binary mask function over the area where inclusion bodies are analyzed; thus $I_{IB}^{\lambda_G}(x,y) = I^{\lambda_G}(x,y) \triangle I_{Mask}^{\lambda_G}(x,y)$, where $\triangle$ represents the bitwise multiplication.

Let us A_{Mask} be defined such as the circular binary mask's total area over the region of interest analyzed of the inclusion bodies, defined by

$$A_{Mask} = \sum_{x,y} I_{Mask}^{\lambda_G}(x,y), \ for \ I_{Mask}^{\lambda_G}(x,y) > 0 . \qquad (12)$$

K-Law nonlinear filter function (K-Law) in pattern recognition is used to analyze and explore the discriminating property quality of each filter [26] over the WSSV inclusion body segmented image $I_{IB}^{\lambda_G}(x,y)$ function.

K-Law filter function is derived by the Fourier transform of the $I_{IB}^{\lambda_G}(x,y)$ function, denoted by

$$I_{IB}^{\lambda_G}(u,v) = \left| I_{IB}^{\lambda_G}(u,v) \right|^k \exp\left[-i\phi(u,v) \right], \ k = 1 . \qquad (13)$$

The K-Law nonlinear filter of $I_{IB}^{\lambda_G}(x,y)$ is applied by the change of value $0 < k < 1$ in (13), where k is the nonlinear strength; thus intermediate values of k permit the variability of filter features [27].

Let $f_w^{\lambda_G}(u,v)$ function be defined by the application of K-Law Fourier related filter, calculated over the $I_{IB}^{\lambda_G}(x,y)$ denoted by

$$f_w^{\lambda_G}(u,v)_k = I_{K-Law}^{\lambda_G}(u,v)_w , \quad 0 < k < 1 \qquad (14)$$

where $k = 0.1$ is used in (13) and u,v are variables in frequency domain.

The *SSF* function can be obtained by

$$SSF\left(f_w^{\lambda_G}(u,v)_k \right) = \begin{cases} 1, & if \ \mathrm{Re}\left(f_w^{\lambda_G}(u,v)_k \right) > 0 \\ 0, & otherwise \end{cases} , \qquad (15)$$

where $f_w^{\lambda_G}(u,v)_k$ is obtained by (14).

Finally, the spectral signature index i^{ss} is developed and can be defined by

$$i^{ss} = \left\{ \frac{SSF\left(f_w^{\lambda_G}(u,v)_k \right)}{\left(A_{Mask} \right)_w} \middle| (u,v) \in \mathbb{C} \right\} , \qquad (16)$$

where *SSF* and A_{Mask} are obtained respectively according by (15) and (12) for every image $f_w^{\lambda}(x,y)$ found in the inclusion bodies filter bank.

E. Spectral signature index classifier block diagram

Figure 8 shows the block diagram of the new spectral technique involved in WSSV inclusion bodies classification by measurement of the spectral signature index i^{ss} over the infected shrimp's tissue patterns.

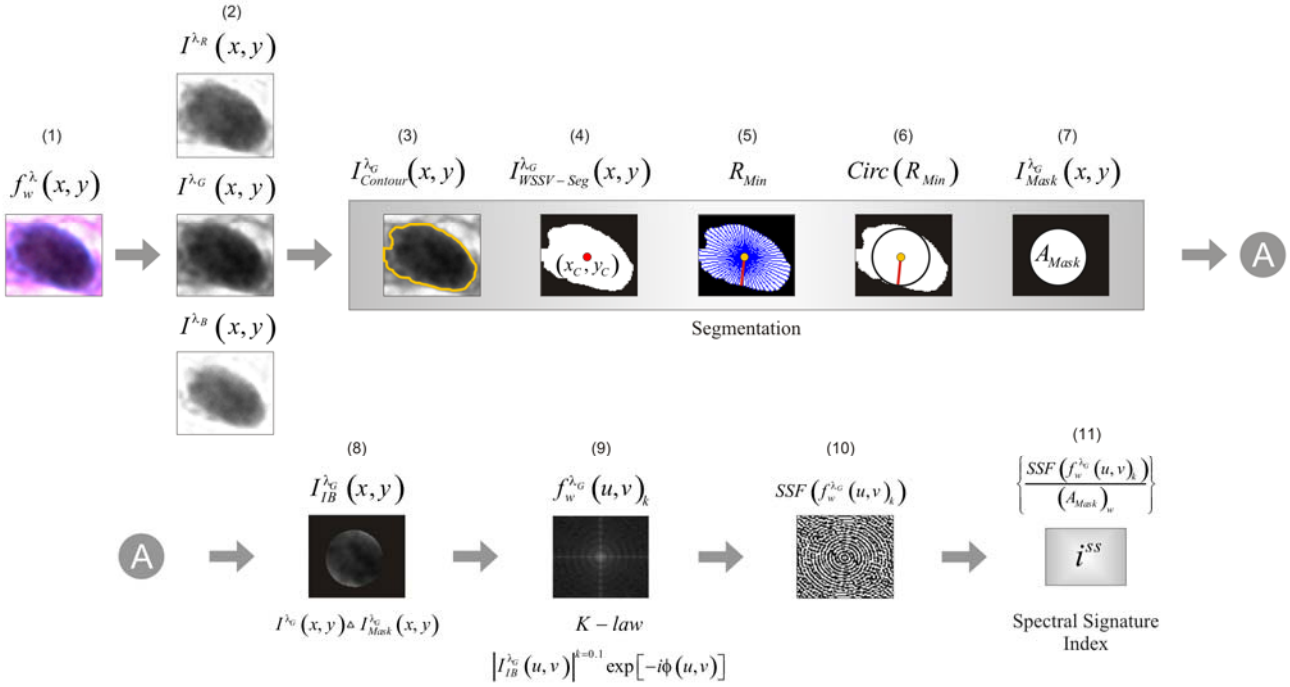


Figure 8. Block diagram to obtain the developed spectral signature index.

This methodology is explained by the following steps: 1) The $f_w^{\lambda}(x,y)$ function is acquired from WSSV inclusion bodies color image sample filter bank; 2) This $f_w^{\lambda}(x,y)$ is divided into its RGB $I^{\lambda_R}(x,y)$, $I^{\lambda_G}(x,y)$ and $I^{\lambda_B}(x,y)$ functions; 3) Segmentation of $I^{\lambda_G}(x,y)$ by using contour "snake" techniques and a morphological operator to get the function $I_{Contour}^{\lambda_G}(x,y)$; 4) Filling the function $I_{Contour}^{\lambda_G}(x,y)$ is obtained the function $I_{WSSV-Seg}^{\lambda_G}(x,y)$, afterwards is calculated the centroid coordinates (x_C, y_C) on this function; 5) From centroid coordinates (x_C, y_C) towards the edge of function $I_{WSSV-Seg}^{\lambda_G}(x,y)$ is obtained the minimum radius R_{Min}; 6) By rotating R_{Min} centered on (x_C, y_C) is obtained the function $Circ(R_{Min})$ used to build the final circular mask; 7) A circular mask function $I_{Mask}^{\lambda_G}(x,y)$ is created inside $I_{WSSV-Seg}^{\lambda_G}(x,y)$ by filling the area of the function $Circ(R_{Min})$, where WSSV inclusion body pattern is analyzed, at same time the area A_{Mask} is calculated from the $I_{Mask}^{\lambda_G}(x,y)$ binary mask function; 8) Using $I_{Mask}^{\lambda_G}(x,y)$ function, segmentation operation is applied over the WSSV inclusion body area $I^{\lambda_G}(x,y)$, where is obtained the $I_{IB}^{\lambda_G}(x,y)$ function; 9) K-Law nonlinear operation is applied in $I_{IB}^{\lambda_G}(x,y)$ function to get the $I_{K-Law}^{\lambda_G}(u,v)_w$ function then it becomes into $f_w^{\lambda_G}(u,v)_k$ function in frequency domain; 10) The frequencies are extracted and analyzed from $f_w^{\lambda_G}(u,v)_k$ by the function SSF and 11) the K-Law spectral signature index i^{ss} is calculated using (16).

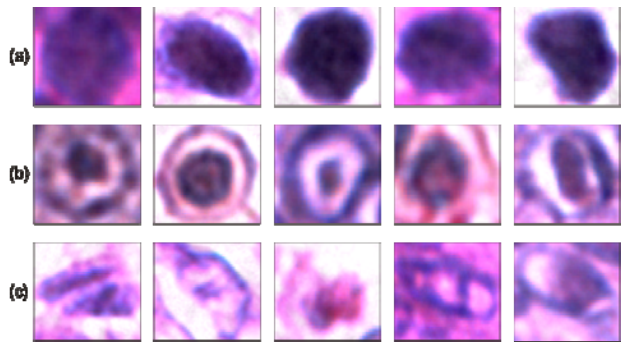


Figure 9. (a) WSSV strong basophilic inclusion bodies, group I; (b) WSSV white halo and chromatin Cowdry type A inclusion bodies, group II; (c) Non-infected tissue particles, group III.

TABLE I. SPECTRAL SIGNATURE INDEX STATISTICAL VALUES

WSSV Group	Signature Index Statistical Behavior			
	$\bar{x}_{i\,ss}$	$\sigma_{i\,ss}$	$1SE$	$2SE$
I	1.3748	0.4817	0.0852	0.1703
II	2.6069	1.8533	0.4953	0.9906
III[a]	159.4229	352.5394	94.2201	188.4402
IV[b]	1.7498	1.2362	0.1823	0.3645

a. Non-infected tissue group particles;

b. Groups I and II analyzed together.

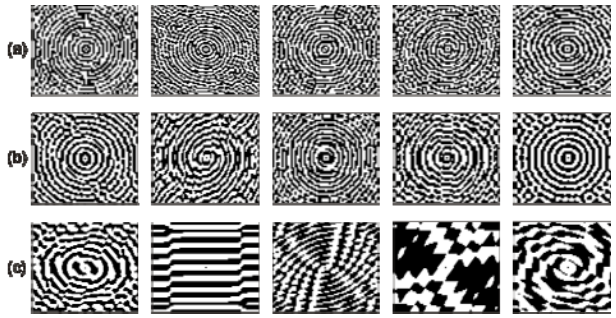


Figure 10. (a) SSF frequencies of group I; (b) SSF frequencies of group II; (c) Non-infected particles SSF frequencies, group III.

III. RESULTS

In order to see if this technique is working with a good performance with the groups of representative inclusion bodies of WSSV with different morphological images, were analyzed with spectral signature index; their images and an additional group of non-infected tissue particles are shown in Figure 9; hence in Table 1 the statistical behavior values of the spectral signature index are shown, including the mean value $\pm 2SE$ (two standard errors).

A set of 168 microscope image field samples were acquired without any additional preprocessing like illumination or contrast correction just the fusion technique developed by [24]; then from this set 870 WSSV inclusion body images were cut, selecting 100 most representative WSSV inclusion bodies to build a filter bank, afterwards every WSSV inclusion body pattern were analyzed to get their frequency measurement and addition to build the biogenic particle cluster by this spectral index.

The great difference in the statistical values of the infected particles with the non infected particles is due to the behavior of WSSV inclusion bodies frequencies in the green channel complex plane, while the non-infected cells do not show similar frequency properties, the infected particles show a well determined frequency signature; in Figure 10 are presented some examples of these frequencies behavior. All the WSSV inclusion bodies were analyzed together; the results shows that the complete inclusion bodies group can be located in well defined fringe $1.3853 \le i^{ss} \le 2.1143$ with $(\pm 2\mathrm{SE})$.

IV. CONCLUSION AND FUTURE WORK

This paper presents a new technique to classify WSSV inclusion bodies from infected shrimp tissue image, based on the analysis of frequencies found in the green channel with K-Law non-lineal filter.

Representative groups of WSSV inclusion bodies from infected shrimp tissues and organs were analyzed. The results show that inclusion bodies are well defined in a clear numerical fringe; thus it can be inferred that whatever analyzed particle with a spectral signature index i_k^{ss} value outside of $1.3853 \leq i^{ss} \leq 2.1143$ range can be considered as non-infected particle.

Future work can be done in the development of automatic WSSV identification system applying this spectral index classifier; however this kind of classifier can be extended by combining complementary frequency analysis techniques by Fourier related filters on the calculation of the spectral signature index to obtain better discriminating results.

Experiments with new tissue samples can be done from others shrimps organs where the virus has a different pattern and its identification is more complex.

Finally, the potential of this signature index can be used to classify other kind of shrimp's viruses and/or other animal, human viruses or biogenic particles.

ACKNOWLEDGMENT

This document is based on work partially supported by CONACYT under Grant No. 102007.

REFERENCES

[1] Bueno-Ibarra M. A., Chávez-Sánchez M. C. and Álvarez-Borrego J., "Development of nonlinear k-law spectral signature index to classify white spot sysndrome virus basophilic inclusion bodies". The First International Conference on Advances in Bioinformatics and Application (BIOINFO 2010), Cancún, México, March 7-13 2010, pp. 30-33.

[2] Lightner D. V. and Redman R. M., "Shrimp diseases and current diagnostic methods", Aquaculture, vol. 164, Issue 1-4, May 1998, pp. 201–220.

[3] Lightner D. V., "A handbook on shrimp pathology and diagnostic procedures for diseases of cultured penaeid shrimp", World Aquaculture Society, Baton Rouge, LA, USA, 1996.

[4] Peinado-Guevara L. I. and López-Meyer M., "Detailed monitoring of white syndrome virus (WSSV) in shrimp commercial ponds in Sinaloa, México by nested PCR", Aquaculture, vol. 251, Issue 1, January 2006, pp. 33–45.

[5] Bell T. A. and Lightner D. V., "A Handbook of Normal Shrimp Histology Special Publication No. 1", World Aquaculture Society, Baton Rouge, LA, USA, 1988, pp. 1–114.

[6] FAO. The State of World Fisheries and Aquaculture. FAO Fisheries and aquaculture Department. Food and Agriculture Organization of the United States. 2008, p. 196.

[7] Holthuis, L.B., FAO species catalogue. Shrimps and prawns of the world. An annotated catalogue of species of interest to fisheries. FAO Fish. Synop., (125) vol. 1: 1980, p. 261.

[8] Briggs M., Funge-Smith S., Subasinghe R. and Philips M., "Introductions and movement of *Penaeus vannamei* and *Penaeus stylirostris* in Asia and the pacific". FAO, Bangkok, 2004, p. 32.

[9] Bondad-Reantaso M.G., Subasinghe R.P., Arthur J.R., Ogawa K., Chinabut S., Adlard R., Tan Z. and Shariff M., "Disease and health management in Asian" Aquaculture. Vet. Parasitol. (132), 2005, pp. 249-272.

[10] Walker Peter J. and Mohan C. V., "Viral disease emergence in shrimp aquaculture: origins, impact and the effectiveness of healt management strategies", Reviews in Aquaculture, vol. 1, February 2009, pp. 125-154.

[11] Chávez-Sánchez M. C. and Montoya-Rodríguez Leobardo, "Enfermedades virales, un reto a la camaronicultura", Primer Foro de Pesca y Acuacultura de las Costas de Chiapas, ECOSUR, Unidad Tapachula, Chapter II, September 2001, pp. 17-24.

[12] OIE. Manual of diagnostic test for aquatic animals. http://www.oie.int/eng/normes/fmanual/A_summry.htm, 2009.

[13] Lightner D. V. and Pantoja Carlos R., " A handbook of penaid shrimp diseases and diagnostic methods (English and Spanish versions.)", A publication in cooperation between the University of Arizona and United States Department of Agriculture (USDA-USAID-CSREES), as part of the Hurricane Mitch Reconstruction Program of Central America (Spanish version), 2001, pp. 48–57.

[14] OIE. Manual of diagsnostic test for aquatic animals. http://www.oie.int/esp/normes/fmanual/A_summry.htm, 2003.

[15] Xu J., Han F. and Zhang X., "Silencing shrimp white spot syndrome virus (WSSV) genes by siRNA", Antiviral Research, vol. 73, Issue 2, Feb 2007, pp. 26–131.

[16] Álvarez-Borrego J. and Chávez-Sánchez M. C., "Detection of IHHN virus in shrimp tissue by digital color correlation" Aquaculture, vol. 194, Issue 1-9, August 2000.

[17] Álvarez-Borrego J. and Fájer-Ávila Emma J., "Identification of platyhelminth parasites of the wild bullsye pufferfish (*Sphoeroides annulatus*, Jenyns, 1853) using invariant digital color correlation", Marine Biology and Oceanography México, vol. 41(1), July 2006), pp. 129-139.

[18] Mouriño-Pérez Rosa R., Álvarez-Borrego J. and Gallardo-Escárate C., "Digital color correlation for the recognition of *Vibrio cholerae 01* in laboratory and environmental samples", Marine Biology and Oceanography México, vol. 41(1), July 2006, pp. 77-86.

[19] Coronel-Beltrán A. and Álvarez-Borrego J., "Comparative analysis between different font types and styles letters using a nonlinear invariant digital correlation", Journal of Modern Optics, vol. 57(1), January 2010, pp. 58-64.

[20] González-Fraga J. A., Kober V., Álvarez-Borrego J. and I. A. Ovseevich, "Pattern recognition of fragmented objects with adaptive correlation filters", Optical Memory & Neural Networks, vol. 15, No. 3, ISSN 1060-99XX, 2006.

[21] Millán, M. S., Campos J., Ferreira C., Yzuel M. J., "Matched filter and phase only filter performance in colour image recognition", Optics Communications, vol. 73(4), 1989, pp. 277-284.

[22] Millán, M. S., Yzuel M. J., Campos J., Ferreira C "Different strategies in optical recognition of polychrormatic images", Applied Optics, vol. 31(14), 1992, pp. 2560-2567.

[23] Bueno-Ibarra M. A., Álvarez-Borrego J., Acho L. and Chávez-Sánchez M. C., "Fast autofocus algorithm for automated microscopes", Optical Engineering, vol. 44(6), 063601-1, 2005.

[24] Bueno-Ibarra M. A., Álvarez-Borrego J., Acho L. and Chávez-Sánchez M. C., "Polychromatic image fusion algorithm and fusion metric for automatized microscopes", Optical Engineering, vol. 44(9), 093201-1, September 2005.

[25] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models.", International Journal of Computer Vision. vol. 1(4), 1987, pp. 321-331.

[26] González-Fraga J.A., Kober V. and Álvarez-Borrego J., "Adaptive SDF filters for pattern recognition", Optical Engineering, vol. 45, 057005, 2006.

[27] Guerreo-Moreno R. E. and Álvarez-Borrego J., "Nonlinear composite filter performance", Optical Engineering, vol. 48(6), 067201-1, June 2009.

An Adaptive Entropic Thresholding Technique for Image Processing and Diagnostic Analysis of Microcirculation Videos

Nazanin Mirshahi, Kayvan Najarian
Dept. of Computer Science
VCU
Richmond, VA, USA
mirshahin@vcu.edu, knajarian@vcu.edu

Sumeyra Demir, Rosalyn Hobson
Dept. of Electrical Engineering
VCU
Richmond, VA, USA
demirsu@vcu.edu, rhobson@vcu.edu

Kevin Ward
Dept. of Emergency Medicine
VCURES
Richmond, VA, USA
krward@vcu.edu

Roya Hakimzadeh
Signal Processing Technologies LLC
Richmond, VA, USA
contact@signalprocessingtechnologies.com

Abstract **– Understanding the functionality of microcirculation is a key factor in the analysis of blood circulatory system. The blood flow distribution changes, based on the physiological effects of disorders. This study presents a method for analysis of microcirculation videos captured from lingual surface of 10 animal subjects. The technique applies advanced image processing methods to stabilize videos, segment microvessels (capillaries and small blood vessels), and estimate the average Functional Capillary Density on 20 consecutive frames for each subject. The algorithm consists of four main parts: pre-processing, video stabilization, entropic-based adaptive local thresholding segmentation and post-processing. The key objective is to quantitatively examine the changes that occur in microcirculation over treatment periods for diseases as well as for the resuscitation process. The designed system will help physicians and medical researchers in diagnostic and therapeutic decision making to determine the sufficiency of resuscitation process and the effect of drug consumption in patients. In particular, the system focuses on minimizing user interaction while improving the accuracy of the analysis. Visual evaluation of the results by medical experts indicates that the technique is capable of identifying 95% of active capillaries and blood vessels in videos.**

Keywords - Microcirculatin, Image processing, multi-resolution, entropic thresholding, Adaptive local thresholding, Lorentz information measure

I. INTRODUCTION

Microcirculation refers to the blood flow in blood vessels less than 100 μm luminal diameter [2]. Changes in microcirculation might be due to numerous diseases and abnormalities in humans. Microcirculatory studies indicate that the small diameter of microvessels (arterioles, capillaries and venules) helps observe changes in blood circulation more evidently compared to large blood vessels. Basically, the rheological properties of blood in capillaries and small blood vessels lead to effective viscosity in those vessels which considerably differentiates the circulation of red blood cells and plasma in microvessels and large blood vessels. The major function of the micro-vascular network is distribution of nutrients, fluid and oxygen throughout tissues in humans [3,4]. As a result, the distributions of microcirculatory network and blood circulation are considered to be key factors in human physiological health [5-8]. Evidence suggests that information regarding the status of microcirculation plays a crucial role in treatment and diagnosis of several diseases such as sepsis, sickle cell, chronic ulcers, diabetes mellitus and hypertension [9-13]. Research and clinical experience show that each of the mentioned diseases uniquely affects characteristics of microcirculation such as structure of capillaries and features of blood flow [14-18]. Hence, investigation of microcirculatory changes has clinical significance in measurement and observation of the changes in response to treatment of microvessels under clinical conditions. Timely detection of such changes potentially helps in taking proper actions which in turns improves the chances of treatment success. A technique to quantitatively assess and monitor these alterations is extremely valuable for further study of such pathological conditions [19]. Particularly, in trauma care, continuous monitoring of microcirculation and measurement of microcirculation indices while resuscitation process helps in determining when to start/stop resuscitation [20-22].

The recent development of videomicroscopy technology has provided effective tools for detection and assessment of tissue perfusion and oxygenation through visualization of microvasculature [23]. Quantitative analysis of microcirculation allows monitoring changes in microvessels that occur due to diseases and other abnormalities. Both visual analysis and use of existing semi-automated video analysis tools are time-consuming and demanding, preventing real-time assessment of microcirculation. This calls for

automated systems to be used for applications including resuscitation.

Two prevalent medical imaging techniques that have been widely used for examining microcirculation during surgery and for other clinical research are Orthogonal Polarization Spectral (OPS) imaging and Side-stream Dark Field (SDF) [24,25]. SDF is superior to OPS in the field of microcirculation study as it improves contrast and lowers surface reflectance compared to OPS [35]. Although advances in hardware systems have played a major role in acquiring knowledge about the physiology and pathology of microvascular function, lack of existing techniques for rapid and accurate processing of microcirculation videos is still an issue. Manual analysis of this information by experts is a complex and time-demanding process which may not be used for real-time assessment of microcirculation; however, an automated system can therapeutically and diagnostically assist physicians and medical researchers.

Several methods have been employed to analyze microcirculation images. A short version of the method proposed in this paper was mentioned in [1]. Dobbe et al. has proposed a semi-automatic, highly accurate method for the analysis of microcirculation [26]. The method applies image stabilization, centerline detection and space time diagram to detect capillaries and small blood vessels. Despite the high accuracy, this method is extremely time-consuming and requires human interaction to produce acceptable results, and therefore, it is not appropriate for real-time applications of microcirculation analysis. The image processing techniques that were proposed in the field of microcirculation are mainly used to process high quality color and/or grayscale retinal images; conversely, the accuracy of the results declines when the same method is applied to other microcirculation images due to their low contrast. Numerous techniques and their combinations have been employed on segmentation of small blood vessels. Pattern recognition-based techniques were used by Staal et al. to analyze two-dimensional color images of retina [27]. Several features of the image are selected and classified to extract image ridges automatically. The main shortcoming of this method is that it is likely to over- and undersegment the vessels. The tracking-based approach described in [28] locates the optic nerve in ocular fundus images. Utilizing fuzzy convergence of the blood vessel, the algorithm uses two features, convergence of vessel network and brightness of the nerve to perform segmentation. Despite its capabilities, the method fails to accurately detect blood vessels where bright lesion regions exist. Vermeer et al. applied a model based approach [29]. The method incorporates Laplace concept, thresholding as well as classification to detect vessels in retinal images. The method requires high levels of human intervention, therefore, is not appropriate for real-time segmentation of microcirculation images. Artificial intelligence methods use prior knowledge for direct segmentation [30].

Most of the vessel segmentation methods that were reviewed earlier in this section are capable of extracting vessels in retinal images; however, lack essential properties to segment microcirculation images [31]. The aim of the proposed study is to stabilize and segment low local contrast microcirculation videos automatically, accurately and in a close to real-time manner in order to aid physicians and clinical researchers in making diagnostic and therapeutic decisions. This algorithm attempts to eliminate human intervention in the precise extraction of small blood vessels. Furthermore, it calculates the diagnostically useful measure of Functional Capillary Density (FCD) for 20 consecutive frames in a microcirculation video [34]. The algorithm segments the image using a modified entropic thresholding technique [33]. Entropy-based methods apply a threshold to the images using entropy of an image or similar information. The rest of thresholding techniques are categorized into five main classes [36]. In histogram shape-based methods, certain parts of image histogram are assessed. Clustering-based methods utilize mixture of two Gaussians to separate foreground and background in an image. Object attribute-based algorithms look at similarities between the original image and its corresponding binary image. Higher order probability distribution is used in thresholding based on spatial methods. Local methods use local image properties to threshold an image. Experiments have shown that for the purpose of thresholding microcirculation images, entropic thresholding techniques yield the most successful outputs.

Section II provides a detailed description of the methodology including preprocessing, video stabilization, segmentation and post-processing. The results of the study on 10 video samples of hemorrhaged and healthy subjects are presented in Section III. Section IV contains the conclusion and discussion of the obtained results. Finally, Section V concludes the paper with future work.

II. METHODOLOGY

The proposed methodology is an extension of [1]. Key modifications were applied to improve the segmentation part. Furthermore, the algorithm was examined on more data samples to evaluate the capabilities of the technique in this paper. The

microcirculation videos used to validate the results of the proposed research study are based on SDF imaging technique, captured by MicroScan hardware [35]. MicroScan is an easy to use instrument that is mainly utilized in various microcirculation observations and analysis. The data samples for this study were acquired from sublingual surface of healthy and hemorrhaged swine subjects. In SDF imaging modality, the light from light emitting diodes is absorbed by hemoglobin, which results in the visibility of flowing cells. Consequently, the walls of the capillaries become visible in the presence of Red Blood Cells (RBCs) [24]. Medical research has proven that lingual surface suffice for the investigation of microcirculation condition in the body as capillaries are adequately superficial for MicroScan. Thus, lingual recordings are considered valid indicators of normality or abnormalities of the microcirculatory network.

A major challenge in the image processing of microcirculation videos are their low resolution and local contrast that complicates the distinction between objects of interest (capillaries and small blood vessels) and frame background. An instance of an original microcirculation frame is provided in Figure 2. Another challenge is the inconsistency of the graylevel intensity in background and blood vessels from one sample to another. The effect of uneven lighting due to the movement of camera and/or subjects results in different levels of intensity in different frames. Therefore, the choice of a single threshold level for an entire frame is not adequate for effective thresholding. The proposed study addresses the mentioned issues by adopting an adaptive entropic thresholding technique.

Prior to being processed, videos are converted to their comprising sets of images or frames; these frames were either processed individually or in combination with other frames as will be described in the rest of the paper. An outline of different steps of the algorithm is provided in Figure 1.

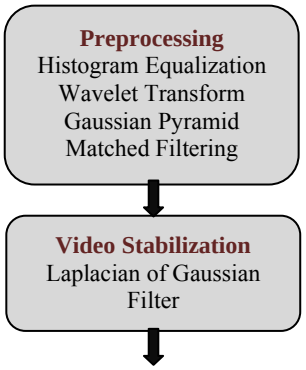


Figure1. Block diagram of the algorithm

A. PREPROCESSING

Preprocessing of microcirculatory images is essential considering the low local contrast of microcirculation images. Preprocessing usually comprise a series of operations to improve the quality of images in order to maximize the difference between image background and objects of interests. In microcirculation images, the intensity of capillaries and small blood vessels are exceptionally close to that of background and tissues. In order to process the images, the first main step is preprocessing.

As the first step, adaptive histogram equalization is applied to the images to help enhance low local contrast of the images. The histogram of an image is a representation for the number of different pixel values in the image. Microcirculation images comprise of a narrow range of intensities. In adaptive histogram equalization, the histogram for various parts of the image is generated and interpolated. Bilinear interpolation eliminates the visibility of the boundary lines that were produced by local histograms. The result of adaptive histogram equalization is a modified image whose histogram is different from that of the original image. In other words, the background appears rather uniform in terms of intensity with a remarkable contrast as compared to the blood vessels and other artifacts in the image.

To further reduce the effects of background noise, wavelet transformation is incorporated in this step. Wavelet transformation decomposes the image into its different frequency contents. Usually, high frequencies represent noise and low frequencies represent details in an image. The image is transformed to wavelet domain and decomposed with mother wavelet of Daubechies 8, level 2. Following that, high frequencies present within the image are filtered. Then the image is reconstructed in the original domain. The noise in the resulting image is much lesser than the input image.

A microcirculation frame usually contains blood vessel in different levels of proximity to the surface of tongue. Representing images in multiple spatial-

frequency domains emphasizes the patterns of blood vessels in different scales that normally can not be seen in the image. The significance of analyzing images at various resolutions is that the objects of different sizes are more visible in different resolution levels. Experiment shows that level 2 and level 3 of Gaussian pyramid separate blood vessels of different sizes, thus making the segmentation a more accurate process. To construct the first level of Gaussian Pyramid, image is filtered using a low-pass filter and then sub-sampled. Low-pass filter presents an equivalent effect as convolving images with a series of Gaussian-like weighting functions followed by sub-sampling. The same procedure is repeated, using the resulting image of the earlier step as an input image to generate the next set of results. The filter operates as convolution of Gaussian blur kernel with the image to eliminate high frequencies components. The pyramid compresses image by making it coarser in each level and reducing the number of bits of precision. The blood vessels become more distinguishable after this step. In this study, for each frame, levels 2 and 3 of Gaussian Pyramid are saved for future analysis.

Matched filter is applied to the image to extract features. In this step, to enhance the edges of blood vessels, matched filter is applied to the image [32]. Matched filter approximates the intensity profile of the image with Gaussian curves. In this study, the function in equation (1) is implemented for the detection of linear anti-parallel pieces of blood vessels. Although the gray-level profile varies for different vessels, similar properties of blood vessels make the mentioned Gaussian function appropriate for this purpose. The function is a two-dimensional kernel that is convolved with the image to sharpen edges of blood vessels.

$$f(x,y) = -e^{(\frac{-x^2}{2\delta^2})} \text{ for } |y| \leq \frac{L}{2} \qquad (1)$$

In (1), L represents the size of the selected slice of the vessel with fixed orientation. The value of L is specified by experiment. The kernel is originally only aligned with y-axis. In order for the kernel to detect vessels in other orientations, the kernel is rotated. The rotation is performed convolving ten 15×16 pixel kernels with the image. The maximum value resulting from each convolution is considered the convolution response of that orientation. This step of the algorithm enhances the edges of blood vessels, while blurring large blood vessels and tissue. An instance of this effect is illustrated in Figure 3.

B. VIDEO STABILIZATION

Recording videos from microcirculation provides an effective tool to visualize the activity of blood vessels and capillaries over a short period of time. This makes video superior to image when analyzing microcirculatory networks, since video contains more information compared to image. Despite the advantages of capturing videos for the study of microcirculation, the motion artifact due to the movement of the handheld camera and that of the subject are obstacles for effective analysis of microcirculation videos. To eliminate the effects of motion artifacts, video stabilization is performed at this step.

The main aim of this step is to calculate the transformation between two consecutive frames in the video. The first step is to compute the first derivatives of the image using Gaussian Gradient filter. Since the image is a function of two variables, $f(x,y)$, the derivatives are computed in both horizontal and vertical direction. The sum of the resulting values of the filter in each direction generates the overall Gaussian gradient for the image. Following that, seven control points are selected in the first frame. The control point should be on the vessels and not on the background in order to be tracked effectively. Laplacian of Gaussian filter is applied to the frame for choosing the most relevant control points. Laplacian of Gaussian filter computes the second order derivatives of the Gaussian function. The filter is used to find regions of rapid change in images such as edges. Using the filter guarantees that the selected points are located on the objects of interest. The control points are the ones that yield to the highest amounts of filter output. Once being picked for the first frame, control points are tracked in the 19 following frames. To track the same points in the following frames, a window of *25x25* pixels is defined around the control point. The points are tracked within a corresponding window of *65x65* size in the following frames. In case any of the point is fallen outside the image due to excessive motion, new control points will be defined using the mentioned method.

The next step is to calculate the cross-correlation coefficients between two successive frames. Cross-correlation coefficients are calculated for tracking the points in the previous step. To identify the control points in the frames except for the first frame, the *65x65* window is scanned to detect the maximum correlated regions in two consecutive frames. The frames are registered according to the maximum correlated sub-areas. Registration helps finding the amount of shift between the two successive frames. In

other words, the distance between the frames is acquired through calculating the cross-correlation value. $d(f_i, f_{i+1})$ represents the distance between the grayscale profile of two succeeding image in which i shows the frame number. This value will be used as a parameter in segmentation. An example of the stabilization result for 10 frames is shown in Figure 4.

C. SEGMENTATION

Image segmentation is performed to partition image into its comprising components. The objective of microcirculation image processing is to separate background from blood vessels and capillaries using grayscale values. Such separations make the analysis of the image an easier task. The outcome of segmentation is a binary image whose background is shown with white pixels and objects of interest with black pixels.

One main class of techniques that is incorporated for image segmentation is thresholding. The main classes of thresholding were mentioned in the introduction part. Depending on the application, one may use global thresholding or local thresholding. Uneven grayscale intensity of the background and the varieties in the intensity of the objects in microcirculation images make the global thresholding inefficient. However, local thresholding smoothly varies across the image and is capable to adapt the threshold value for different parts of the image. This study adaptively selects windows in the image based on Lorentz Information Measure (LIM) [37]. Following that, for the thresholding of each sub-image, the algorithm implements an extension of the entropic thresholding technique proposed in [33].

Adaptive local thresholding successfully reduces the issue with uneven background intensity by partitioning the image into windows of variable sizes based on LIM value. If the image F is the gray level image and $f(x,y)$ is defined as the intensity image, the amount of information in the image known as Picture Information Measure (PIM) is calculated by:

$$PIM(f) = \sum_{i=0}^{F-1} h(i) - \max h(i) \qquad (2)$$

PIM shows the minimal graylevel variation if $f(x,y)$ is converted to a constant grayscale image. $h(i)$ shows the graylevel histogram of the image for $h: \{0,1,...,F-1\} \rightarrow N$. If the image comprises of only one graylevel, $PIM(f) = 0$, while $PIM(f) = \max$ occurs when the graylevel intensity of the image is uniformly distributed.

The normalized form of (1) is defined as:

$$NPIM(f) = PIM(f) / N(f) \qquad (3)$$

where $N(f)$ is the number of pixels in the image. $NPIM(f)$ can be calculated for any sub-image in the image, indicating the amount of information in the given sub-image. An experimental cutoff value for $NPIM(f)$ is chosen in order to adaptively adjust the sub-image size for thresholding. The values greater than the cutoff value ensure that the sub-images contain both background and objects of interest.

The superiority of thresholding method in the proposed algorithm is that it considers flow information in addition to local property and intensity information. The technique is based on graylevel spatial correlation histogram of the image. The aim is to maintain the spatial structure of the image using pixel neighborhood property. In an image F with graylevel intensity of $f(x,y)$ in which (x,y) represents the coordination of a pixel, let $Q \times R$ be the number of pixels. $g(x,y)$ is defined as the number of pixels in $N \times N$ neighborhood of pixel (x,y) within ζ distance of the pixel. The ζ value of 2 and neighborhood of 3×3 ($N = 3$) were chosen empirically for the purpose of this study.

The flow factor used in this part was calculated in the stabilization section. For every frame, the summation of the differences between each consecutive pair frames in 10 preceding frames is computed using the following equation:

$$Sd(f) = \sum_{n=i-10}^{n=i} d(f_n, f_{n+1}) \qquad (4)$$

Three parameters of flow, neighborhood vicinity and pixel intensity information are used in equation (5) to calculate the graylevel spatial correlation histogram of the image:

$$h(k,m,D) = prob(f(x,y) = k, g(x,y) = m, Sd(f) = D) \qquad (5)$$

Equation (5) determines the probability $h(k,m,D)$ in which $f(x,y) = k$, $g(x,y) = m$ and $Sd(f) = D$ where $0 \le k \le 255$, $1 \le m \le 9$ and $1 \le D \le 2$.

According to the principle of entropy, noise and edge produce more information than background and

objects. In order to emphasize the effect of *m* as a key distinction factor, the nonlinear function in equation (6) is multiplied with entropy function.

$$W(m,N) = (1+e^{\frac{-9m}{N\times N}}) / (1-e^{\frac{-9m}{N\times N}}) \qquad (6)$$

In equation (6), N is the selected neighborhood of a pixel and *m* is the number of neighbor pixels within ζ distance of the pixel, $m = \{1,2,3,...,N\text{x}N\}$.

In the next step, threshold *T* is calculated; *T* is $0 < T < l-1$. It segments the image into object and background, represented by *O* and *B* respectively. In order to calculate *T*, the second order entropy of image and background is calculated using equations (7) and (8).

$$H_B^{(2)}(T,N)_{=} - \sum_{k=T+1}^{255}\sum_{m=1}^{N\times N} \frac{P(k,m,D)}{P_B(T)} \ln(\frac{P(k,m,D)}{P_B(T)}) W(m,N) \qquad (7)$$

$$H_O^{(2)}(T,N)_{=} - \sum_{k=0}^{T}\sum_{m=1}^{N\times N} \frac{P(k,m,D)}{P_O(T)} \ln(\frac{P(k,m,D)}{P_O(T)}) W(m,N) \qquad (8)$$

In the equations of second order entropy, $P(k,m,D)$ is the normalized form of $h(k,m,D)$.

The optimal threshold is calculated as the total of equations (7) and (8). The optimal threshold is calculated using:

$$H(T,N) = H_O^{(2)}(T,N) + H_B^{(2)}(T,N) \qquad (9)$$

T is obtained by computing a value that maximizes *H(T,N).* Experimental evidence has shown that as a result of noise factors, the obtained value is not the optimal threshold. In order to eliminate this issue, the median of ten maximum values of *H(T,N)* is selected to be the optimal threshold.

The final segmentation process employs the information acquired through calculation of Lorentz Information Measure to threshold the image using the mentioned entropic thresholding technique. As mentioned in the stabilization part, 20 frames are stabilized for each level of the Gaussian Pyramid. The output of the stabilization is a set of stabilized frames that demonstrate active blood vessels. In many instances, stabilization distorts the edges of the frames to effectively smooth out the video. In order to eliminate the possible effect of the stabilization, 15 pixels from top, bottom, right and left of the images are removed. Following that, the intensity values of each pixel coordinate in 20 stabilized frames are arithmetically averaged. The result of this step is the input for the main segmentation part.

The size of microcirculation images of the study after removing a 15 pixel frame from the image becomes *450x690* pixels. The original window size of *60x60* pixels is chosen empirically to divide the image into sub-images for thresholding. Thresholding starts with the original window size from the top left of the image. The image is partitioned into a window of size *60x60* pixels and $NPIM(f)$ is calculated for the sub-image. If the $NPIM(f)$ value was greater than 0.97, the sub-image is thresholded using the mentioned entropic thresholding technique. The limit value for $NPIM(f)$ was selected experimentally. If the $NPIM(f)$ value is less than the limit, the window size adaptively grows to *120x120*, twice as much as the original window in direction of *x* and *y*. The sub-image is then thresholded using the proposed entropic thresholding technique. The same process of thresholding the sub-images is repeated for the entire image. The output of thresholding is a binary image in which blood vessels and capillaries are represented by black pixels and the background with white pixels.

Despite the success of pre-processing to reduce the effect of image artifacts, the result of this step might still contain tissue and other artifacts in shape of scattered small objects. One solution to eliminate the effect of artifacts is to apply morphological operations to the image. Morphology performs mathematical techniques on images to analyze and process geometrical shapes. In this step, the objects in the binary image are labeled. The size, width and length of the objects are acquired and compared to the user defined ones. The objects with values out of the defined range are removed. Such operation clears the image from isolated pixels with width and length less than 4 and 10 pixels as well as large vessels with width of greater than 30 pixels. The result of this step is a binary image with less noise.

D. POST-PROCESSING

Post-processing refers to a combination of techniques for generating the final results of the algorithm. In the previous steps, levels 2 and 3 of Gaussian pyramid were generated for 20 frames. Following that, the frames were stabilized, averaged

and segmented for each level. Accordingly, two sets of images were acquired.

Different levels of Gaussian pyramid provide different resolution of the same image. Since different blood vessels might be better visible in each of the levels, there is a need to combine the results of the two levels to construct the final results. The union of points identified as capillaries and blood vessels in each set forms the ultimate results of the algorithm. The final result is an image that shows the active blood vessels.

III. RESULTS

To verify the effectives of the proposed algorithm, it was tested on 10 microcirculation video samples. Five of the subjects were hemorrhaged animals and five were healthy animal subjects. Equal number of normal and abnormal subjects helps examining the major difference in statistical analysis of the results. The videos were captured with the rate of 30 frames per second. The original size of a video frame is 480x720 pixels. The sample data were provided by Virginia Commonwealth University Reanimation Engineering Shock Center.

The algorithm was applied to the first 20 frames of each video. The results of different steps of the algorithm for a healthy case are illustrated in Figures 2-5, while Figures 6-9 show the results for a hemorrhaged case. MATLAB® programming language was used to develop codes to examine the validity of the proposed algorithm and to generate experimental results. With respect to time complexity, the algorithm takes an average of 15 minutes to run on a 2.40 GHz computer with 3 GB of RAM. The evaluation of results is performed through visual inspection of medical experts. The inspection has shown that the accuracy of algorithm in extracting active blood vessels and capillaries is 95% on average.

The measure of FCD was calculated for the sample data. The FCD value results are listed in Table 1. FCD is the area of the segmented capillaries in an image divided by the area of the image [33]. In each averaged frame, the total number of black pixels is divided by the size of the image to obtain FCD value. The result shows that the algorithm can successfully distinguish between normal and abnormal cases based on a simple statistical analysis.

IV. CONCLUSION

The proposed method is a fully automated approach for image processing of microcirculation videos. The algorithm incorporates a novel thresholding technique that considers flow information to be a key factor in calculation of entropy. Furthermore, it adjusts the threshold level locally based on image information using Lorentz Information Measure. The algorithm looks at two levels of Gaussian Pyramid resolutions to acquire a true estimate of active blood vessels in a video. The technique is capable of distinguishing between the healthy and hemorrhaged subjects in the 10 studied samples using Functional Capillary Density. Visual evaluation of the results shows 95% accuracy in blood vessel detection. The designed technique can potentially assist physicians and medical researchers in making diagnostic decisions.

V. FUTURE WORK

As future work, an extension of the current approach will combine multi-resolution concept with multi-thresholding to improve the segmentation results while reducing the false positives. Other diagnostically useful measures such as Perfused Vessel Density (PVD), Proportion of Perfused vessels (PPV) and Microvascular Flow Index (MFI) will be calculated using the proposed algorithm. A larger dataset will be acquired and the algorithm will be tested and validated on the new dataset. The stabilization technique will be improved and combined with other registration techniques. The results will be validated using the available semi-automated commercial software tools such as Vascular Analysis Commercial software tools by medical experts. Statistical analysis will be performed for further evaluation of the results.

Table1. FCD for five healthy and five Hemorrhaged subjects

	Healthy	Hemorrhaged
Case 1	0.12	0.09
Case 2	0.15	0.08
Case 3	0.14	0.05
Case 4	0.10	0.05
Case 5	0.12	0.07

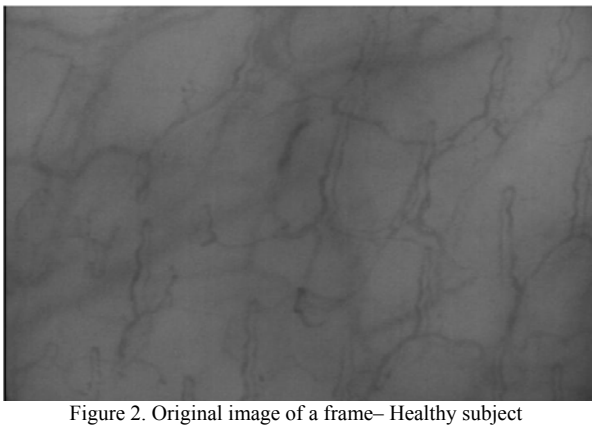


Figure 2. Original image of a frame– Healthy subject

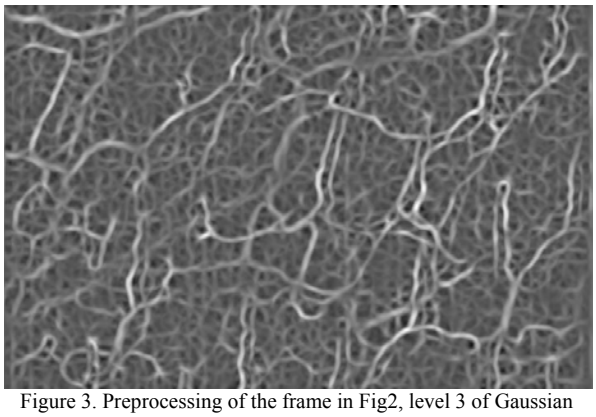


Figure 3. Preprocessing of the frame in Fig2, level 3 of Gaussian Pyramid – Healthy subject

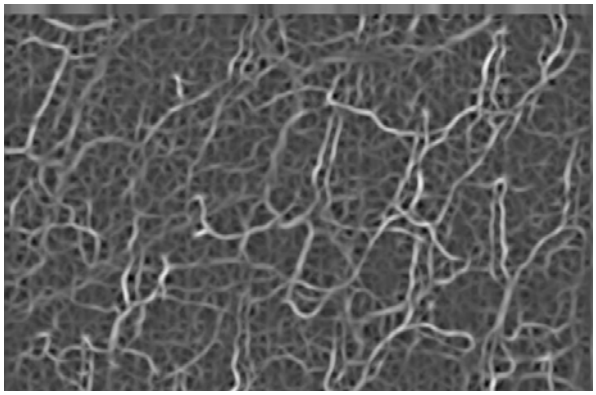


Figure 4. Stabilization results of 10 consecutive frames (Fig2 as the first frame), level 3 of Gaussian Pyramid - Healthy subject

Figure 5.Postprocessing results of 20 frames – Healthy subject

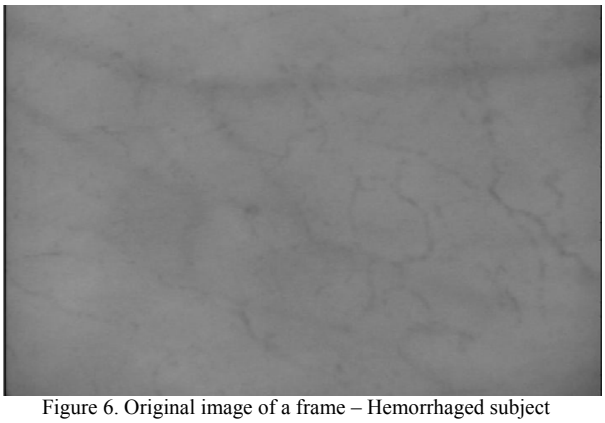


Figure 6. Original image of a frame – Hemorrhaged subject

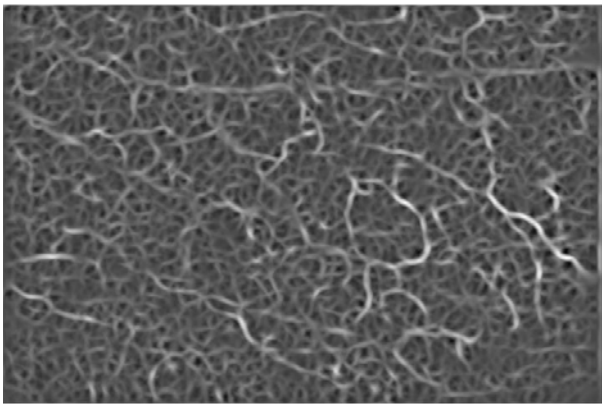


Figure 7. Preprocessing of the frame in Fig6, level 3 of Gaussian Pyramid – Hemorrhaged subject

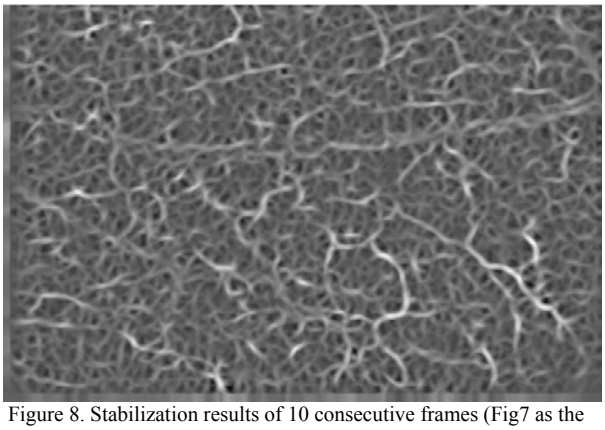


Figure 8. Stabilization results of 10 consecutive frames (Fig7 as the first frame), level 3 of Gaussian Pyramid - Hemorrhaged subject

Figure 9.Postprocessing results of 20 frames - Hemorrhaged subject

ACKNOWLEDGEMENT

The dataset used in this study was provided by Virginia Commonwealth University, Department of Emergency Medicine and Virginia Commonwealth University Reanimation Engineering Shock Center (VCURES).

REFERENCES

[1] N. Mirshahi, S.Demir, K. Ward, R. Hobson, R. Hakimzadeh, K. Najarian, "A Multi-resolution Entropic-Based Image Processing Technique for Diagnostic Analysis of Microcirculation Videos," *biosciencesworld, pp.64-69,* The First International Conference on Biosciences, 2010

[2] R. F. Tuma, W. N. Duran, K. Ley. *HANDBOOK OF PHYSIOLOGY MICROCIRCULATIO,* Elsevier Science, USA, 2008.

[3] AR. Pries, D. Neuhaus and P. Gaehtgens, "Blood viscosity in tube flow: dependence on diameter and hematocrit." *Am J Physiol,* 263: H1770-H1778, 1992.

[4]A. Krogh, *Anatomy and physiology of capillaries*. New Haven Connecticut: Yale University Press, 1929.

[5]A. Krog, "Supply of oxygen to the tissues and the regulation of the capillary circulation." *J physiol (Lond)* 52:457-474, 1919.

[6] AG. Tsai, PC. Johnson and M. Intaglietta, "Oxygen gradients in the microcirculation", *Physiol Rev* 83:993-963, 2003.

[7] CC. Michel and EE. Curry, "Microvascular permeability", *Physiol Rev* 79: 703-761, 1999.

[8] U. Yuan, WM. Chilian, HJ. Granger and DC. Zaeija, "Flow modulates coronary venular permeability by a nitric oxide-related mechanism." *Am J Physiol* 263:H641-H646, 1992.

[9] RM. Bateman, MD. Sharpe, and CG. Ellis, "Bench-to-bedside review: microvascular dysfunction in sepsis: hemodynamics, oxygen transport and nitric oxide" *Crit Care Med 7*: 359–373, 2003.

[10] RP. Hebbel, R. Osarogiagbon, D. Kaul, "The endothelial biology of sickle cell disease; inflammation and chronic vasculopathy" *Microcirculation* 11:129-151, 2004.

[11] MJ. Stuart, and RL. Nagel, "Sickle cell disease", *The Lancent* 364:1343-1360, 2004.

[12] BI. Levy, G. Ambrosio, AR. Pries, and HA. Struijker-Boudier. "Microcirculation in hypertension: a new target for treatment?" *Circulation* 104:735–740, 2001.

[13] C. Verdant, and D. De Backer, "How monitoring of the microcirculation may help us at the bedside", *Curr Opin Crit Care*, 11(3):240-244, 2005.

[14] KA. Nath, ZS. Katusic, and MT. Gladwin, "The perfusion paradox and instability in sickle cell disease", *Microcirculation,* 11:179-193, 2004.

[15] DK. Kaul, and ME. Farby, "In vivo studies of sickle red blood cells", *Microcirculation*: 11:153-156, 2004.

[16] RM. Touyz, "Intracellular Mechanisms Involved in Vascular Remodeling of Resistance Arteries in Hypertension: Role of Angiotensin II", *Exp Physiol,* 90: 499-455, 2005.

[17] EH. Serne, RO. Gans, JC. ter Maaten, GJ. Tangelder, AJ. Donker, and CD. Stehouwer, "Impaired skin capillary recruitment in essential hypertension is caused by both functional and structural capillary rarefaction", *Hypertension*: 38:238-242, 2001.

[18] MA. Creager, TF. Luscher, F. Cosentino, and JA. Bechkman, "Diabetes and Vascular disease: pathophysiology, clinical consequences, and medical therapy: part I", *Circulation* 108:1527-1532, 2003.

[19] O. Genzel-Boroviczeny, J. Strotgen, A. G. Harris, K. Messmer, and F. Christ, "Orthogonal polarization spectral imaging (OPS): A novel method to measure the microcirculation in term and preterm infants transcutaneously", *Pediatr Res* 51:386–391, 2002.

[20] Y. Sakr, MJ. Dubois, D. De Backer, J. Creteur, JL. Vincent, "Persistent microcirculatory alterations are associated with organ failure and death in patients with septic shock", *Crit Care Med*, 32:1825-1831, 2004.

[21] PE. Spronk, C. Ince, MJ. Gardien, KR. Mathura, HM. Oudemans-van Straaten, and DF. Zandstra, "Nitroglycerin in septic shock after intravascular volume resuscitation", *Lancet*, 360:1395-1396, 2002.

[22] M. Fries, M. H. Weil, Y. Chang,, C. Castillo, and W. Tang, "Microcirculation during cardiac arrest and resuscitation", *Crit Care Med* 34: 454–457, 2006.

[23] E. Chaigneau, M. Oheim, E. Audinat, and S. Charpak, *Two Photon imaging of capillary blood flow in olfactory bulb glomeruli.* Proc Natl Acad Sci, USA, 100:13081-13087, 2003.

[24] V. Cerný, Z. Turek, and R. Pařízková, "Orthogonal polarization spectral imaging: a review", *Physiol. Res.* 56, 2007.

[25] C. Ince, "The microcirculation is the motor of sepsis", *Critical Care,* 9(suppl 4):S13-S19, 2005.

[26] J. G. G. Dobbe, G. J. Streekstra, B. Atasever, R. van Zijderveld and C. Ince, "The measurement of functional microcirculatory density and velocity distributions using automated image analysis", *Med Biol Eng Comput,* 46(7): 659–670, 2008.

[27] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. V. Ginneken, "Ridge-Based vessel segmentation in color images of the retina", *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[28] A. Hoover, and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels", *IEEE Tran. on Medical Imaging*, Vol. 22, No. 8, p. 951-958, 2003.

[29] K.A. Vermeer, F.M. Vos, H.G. Lemij, and A.M. Vossepoel, "A model based method for retinal blood vessel detection", Comput. Biol. Med., in press, DOI: 10.1016/S0010-4825(03)00055-6, 2003.

[30] U. Rost, H. Munkel, and C. E. Liedtke, "A knowledge based system for the configuration of image processing algorithms", *Fachtagung Informations und Mikrosystem Technik*, 1998.

[31] C. Kirbas, and F. Quek, "Vessel Extraction Techniques and Algorithms : A Survey" , *Third IEEE Symposium on BioInformatics and BioEngineering*, 2003.

[32] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two dimensional matched filters:' *IEEE Trans. Medical imaging*, vol. 8, no. 3, 1989.

[33] Y. Xiao, Z. Cao, and T. Zhang, "Entropic Thresholding Based on Gray Level Spatial Correlation Histogram", *IEEE Trans. on Wireless Communications 7(1): 334-342*, 2008.

[34] D. De Backer, S. Hollenberg, C. Boerma, P. Goedhart, G. Büchele, G. Ospina-Tascon, I. Dobbe, C. Ince, "How to evaluate the microcirculation: report of a round table condference". *Critical Care*, 11:R101, 2007.

[35] D. Johnson, "Introducing the MicroScan". MicroVision Medical. July 1, 2010 <http://www.microvisionmedical.com/mvm_docs/MicroScan.pdf>.

[36] S Mehmet, S Bulent, "Survey over image thresholding techniques and quantitative performance evaluation" *Journal of Electronic Imaging*, 2004.

[37] S.K. Chang, "Principle of Pictorial Information Systems Design", *Prentice-Hall, Inc.* 1989.

Experiences and Preferences of Patients Regarding a Rheumatology Interactive Health Communication Application: A qualitative Study

Rosalie van der Vaart[1], Constance H. C. Drossaert[1], Erik Taal[1], Mart A. F. J. van de Laar[1,2]
[1] IBR: Research Institute for Social Sciences and Technology, University of Twente, Enschede, The Netherlands
[2] Department of Rheumatology, Medisch Spectrum Twente
Enschede, The Netherlands
r.vandervaart@utwente.nl, c.h.c.drossaert@utwente.nl,
e.taal@utwente.nl, m.a.f.j.vandelaar@utwente.nl

***Abstract* – Interactive Health Communication Applications (IHCAs) can make a valuable contribution to rheumatological care. The development of online health applications is moving quickly, and positive results have been shown. Yet solid research on use and acceptance of different information, communication and participation tools by patients is still lacking. In this qualitative study, we examined the health-related internet use of patients with rheumatic diseases, their motives for using or not using certain applications, and their needs and preferences with regard to a rheumatology IHCA. We conducted semi-structured individual interviews with eighteen patients, who were selected from a hospital's patient panel. Participants were diagnosed with eight different forms of rheumatism and their mean age was 50.7 years. The interviews were analyzed by two independent researchers. Results show that the applications most preferred by participants were: information provision on both medical and support topics, online communication with their doctor and insight in their personal health records. Patient support groups were less valued, as were participation tools such as symptom monitoring and online exercise programs. Patients reported clear preferences and pre-conditions that should be fulfilled in order for them to use the applications. A large discrepancy was found between patients' current use and their future preferences with respect to information about care and support, online access to medical health records and having online contact with their doctor. In conclusion, patients see great value in an IHCA provided by their own hospital, since it could increase reliability of the provided information, and would give them the confidence to use the application. Overall a rheumatology IHCA should contain communication and participation tools, both linked to the hospital, and information about disease, care and practical support. The reported motives and preconditions of the respondents outline key issues which should guide the development of an online application.**

Keywords – IHCA, rheumatism, patients, preferences.

I. INTRODUCTION

The internet is making an increasing impact on today's health care and the expectations about the effects of internet applications in health care are high. First, internet applications could support the growing need for health care resulting from both our aging population and the increasing number of people who suffer from one or more chronic diseases. Second, internet applications offer the opportunity to extend the patients' role in delivering health care [2][3]. Accordingly, such applications could support the transformation of the patient from passive receiver of care into an active participant in the management of one's illness, which is considered highly desirable in chronic health care [4][5][6].

Presently, patients with various chronic diseases can go online to find information, self-tests, and self-help tools or to get in contact with peer patients. In addition, health care organizations and health care providers are increasingly developing their own web applications for their patients. These applications sometimes provide – besides the above mentioned tools – opportunities for online contact with health professionals and/or access to patients' personal health records. Overall, three main categories of online health care applications can be distinguished: (1) information, (2) communication and (3) participation. Information applications mostly hold the provision of disease information and care information. Communication applications concern facilities for communication with peers or with health professionals. Participation – a broad area – concerns applications aimed at symptom monitoring, self-management and access to medical health records.

A. Interactive Health Communication Applications

Interactive Health Communication Applications (IHCAs) are operational software programs which combine the provision of health information with at least one of the above-mentioned communication or participation applications. Patients with chronic diseases, such as rheumatism, can benefit particularly from IHCAs, since such patients are often considered to be on an 'illness journey': as patients progress through their journey, they might have different needs with respect to information, self-management and support needs [7][8][9]. An IHCA has the potential to meet these multiple needs because it provides a wide range of information, communication and participation tools. It is accessible independent of time and place, and its content can be patient tailored – which also supports the patients' personal illness journey [3][10]. Moreover, the information

can be presented in accessible formats, such as video and audio clips, and graphics. Above all, two recent systematic reviews suggest that chronic health care IHCAs are effective in improving knowledge, perceived social support, and health behavior for various kinds of chronic diseases, as asthma, diabetes and heart failure [11][12].

B. Lack of supply and acceptance

Despite these benefits, online applications for rheumatic patients remain scarce. Murray's systematic review included 24 randomized controlled trials on IHCAs, but no rheumatology application [11]. Another systematic review of online self-management systems by Solomon also did not include a rheumatology application in any of the 28 articles reviewed [13]. Our own literature search revealed only one study about a website for patients with rheumatic diseases that combined information, patient-provider communication and health assessment tools [14]. Other existing online rheumatism applications are single applications focusing mainly on participation, such as symptom monitoring, exercise support, or overall self-management [15][16][17].

Whereas IHCAs thus seem to be effective, it is still unclear which combination of tools contributes to these successes [18]. Moreover, not all applications on an IHCA are equally well used [19]. It seems that simply developing and implementing online applications does not suffice. Roughly, the often experienced lack of acceptation seems to be related to the patient on the one side, and to the technology on the other side [20][21]. Usability problems often occur; applications are not being developed patient-centered and are not being tested by users before implementation. Furthermore, it is often reported in studies that patients experience an overload of websites containing information and support possibilities. Yet internet applications are often not initiated from the demand side of the patient, so they do not meet patients' needs. In sum, often it is not known whether the services offered on the internet are services that patients actually desire. Furthermore, acceptance problems are often explainable by patients' existing (negative) attitudes towards innovations [22]. Many people experience doubts on reliability when it comes to health related technology, for the large amount of supply causes confusion on what sources are trustworthy. Furthermore, privacy issues are of large concern to patients when it comes to private health information that is communicated via The Web. Overall, it is important to carefully match the applications on an IHCA to the needs of the patients, so that the offer is patient-centered and actually valuable for them [1].

C. Interview study

In summary, while studies on the needs of patients regarding online applications have been conducted for other chronic diseases [8], within rheumatism there remains a gap in this kind of knowledge. The aim of this study was to perform a needs assessment among patients with rheumatic diseases regarding an IHCA. Our study focused on four questions: (1) Which (information, communication and participation) support applications do rheumatism patients already use on the internet? (2) What are their attitudes about available online support applications? (3) What are their preferences and demands for a rheumatology IHCA? And most importantly, (4) What are their reasons for preferring or not preferring certain applications? This paper will give an expansion on earlier presented work [1] and describes an overview of the methods used in our study, the results that were found in the three main categories of applications and a discussion on each category, including study limitations and a conclusion.

II. METHODS

A descriptive qualitative design was used, since this study was explorative. We preferred the use of individual semi-structured interviews to get the best understanding of patients' experiences, needs, motives and preferences for a selection of widely used internet applications.

A. Selection of participants

Participants were selected from an existing patient panel, which was initiated in cooperation between the University of Twente and Twente's largest clinical hospital. Patients registered on this panel are willing to volunteer in rheumatology research. The criteria for patient participation for the present study were: willing to participate in interviews, contactable by e-mail and not older than 60 years. The interviews took place at the university or at people's homes, at each participant's choice. In total, 18 interviews were conducted, after which data saturation was reached; meaning that no more new information of value was obtained [23, 24].

B. Interview structure

Each interview started off broadly, by asking participants about their internet use. Both general internet use and health and rheumatism related internet use was asked about. Subsequently, participants were asked to reflect freely about their ideas and preferences for a rheumatology IHCA. The interview continued by discussing 7 types of widely-used applications within the three main categories of online health support: information, communication and participation. For each type, a prototype card was made which showed representative examples of existing internet applications and websites. The participants were asked about their current use, their needs and their attitudes regarding these applications. Also important were their motives for use or nonuse and their preferences for the applications. The 7 illustrated cards showed: (1) information about disease and treatment; (2) information about care and support; (3) peer support groups; (4) e-consultations via e-mail or online chat; (5) symptom monitoring (scoring of variables such as pain, swollen joints, mood and activity through which is visualized in graphs); (6) exercise programs; and (7) access to medical health records (the ability to give patients access to their own medical files, with information about their diagnosis,

treatment plan and latest lab results). The interviews took one to two hours, depending on the patient. The interviews were audiotaped, provided patients had given permission beforehand.

C. Data-analysis

The audiotapes of the interviews were transcribed verbatim. Current use and needs were extracted, and citations about attitudes and motives for use, nonuse and preferences or pre-conditions were selected and coded into categories by two independent researchers (RvdV, CHCD). The final categories were defined by consensus between the two researchers. Next, the first researcher examined the raw data again to ensure the robustness of the analytical process and to confirm that all the data were indeed reflected in the coding [24]. During this process, only the participant numbers were used to protect the anonymity of the participants.

III. RESULTS

This section gives an overview of participants' current (health related) internet use and their attitudes towards future use of applications on a rheumatology IHCA.

A. Characteristics and internet use

Eighteen participants were interviewed: five male and thirteen female, with a mean age of 50.7 years (SD = 9.27). Participants interviewed had been diagnosed with eight different forms of rheumatic diseases: more than half of the participants were diagnosed with rheumatoid arthritis (n = 10), two with osteoarthritis. The remaining participants were all diagnosed with a less common rheumatic disease. All participants owned a computer and had home access to the internet. They used the internet on a regular basis, generally for several hours a day. The internet was mainly used for e-mail, obtaining information, purchasing goods and banking. All the participants reported that they had used the internet for health-related purposes, usually to search for information.

B. Utilization of and attitudes toward health related internet applications

Overall participants saw great value in an IHCA provided by their own hospital. They reported it would lower barriers such as unreliability of information, and would give them the confidence to use the IHCA. When asked an open-ended question about which applications participants would like to find and use on a rheumatology IHCA, participants mentioned various topics. Most frequently mentioned were: information on the latest developments in treatment and medication, insight into hospital procedures, and tips to cope with troubles in daily life (e.g., at work, when shopping or doing household chores). All these topics were covered in the themes that were discussed using the prototype cards. Table 1 shows an outline of participants' current use and needs, and their motives for use or nonuse on the 7 themes. Table 2 shows an outline of the preferences that patients reported for each support tool. Both of these tables are being extensively clarified in this section, using participants' quotes.

The applications most preferred by participants were information provision on both medical and support topics, online communication with the doctor and insight in their medical health record. Patient support groups were less preferred, as were participation tools such as symptom monitoring and online exercise programs. What stands out is the discrepancy between current use and future preferences on information about care and support, online communication with the doctor and access to medical health records.

TABLE I. CURRENT USE, NEEDS AND MOTIVES OF PARTICIPANTS TOWARDS ONLINE APPLICATIONS (N = 18)

Application	Use[a]	Needs[a]	Motives pro	Motives con
Information about disease and treatment	high	high	- easy and fast - can read what one wants - can read it when one wants	- information overflow - can be unreliable - confrontational/can cause worry - already has all the necessary information - gets information otherwise
Information about care and support	moderate	high	- structured and complete - overview - helpful in decision-making - good reference tool	- no additional care necessary - current health professionals recommend or refer to supplementary care
Patient support groups	moderate	moderate	- recognition - support in coping - giving and receiving advice - anonymous	- unreliable information/advice - complaining people - confronting - impersonal - not wanting to spend much time on the

				disease
Ask your doctor	low	high	- accessible and easy - reliable - enables time to write down questions and (re)read answers - could save visit to doctor	- non synchronous communication - waiting time for a response
Symptom monitoring	low	moderate	- better disease insight for one self and the doctor - new and fun to try - shows patterns over time	- confronting - time consuming - gets one too focused on pain and signs
Exercise programs	moderate	moderate	- help maintain self-respect - comfortable to exercise and get support at home	- no self-discipline - already exercises by themselves/at a therapist - doubtful accuracy and safety
Access to medical health record	low	high	- more involvement in treatment - overview of appointments - overview of previous and current (lab)results	- too difficult to understand

a. Low: < 6 participants reacted positively; Moderate: 6 - 12 participants reacted positively; High: > 12 participants reacted positively

TABLE II. PREFERENCES AND PRE-CONDITIONS OF PARTICIPANTS FOR ONLINE APPLICATIONS (N = 18)

Applications	Preferences and Pre-conditions
Information about disease and treatment	Information on three topics: - disease (diagnosis, symptoms, heredity) - treatment (medication, therapies, protocols) - coping (psychological, social, tips and tricks)
Information about care and support	Information on two topics: - medical care (job description, specializations, hospital procedures) - practical support (tools, insurances, facilities for e.g. work, housekeeping)
Patient support groups	- positive topics; tips & tricks - divers target groups - good control and protection on posts and privacy
Ask your doctor	- valuable extension to current care but no replacement - contact with own health professional - use for minor/non-urgent questions - quick handling of e-mails
Symptom monitoring	- tele-monitoring by doctor - use in consult and treatment - overview in graphs
Exercise programs	- solution to self-discipline barrier - safe exercises - online coach
Access to medical health record	- clear information and instructions - good protection

1) Information about disease and treatment

Every participant reported having searched for information on rheumatism on the internet. Most of the topics patients had searched for were related to medication, such as user instructions, side-effects and the development of new medications. Participants also went online when they felt pain, when they had doubts about their symptoms or when they had noticed new symptoms. Furthermore, the internet was used to gather information after participants had been given their diagnosis and when they heard or read something interesting. A final reason to search the internet was when a person had forgotten to ask the doctor something. The greatest reported benefits of online information were that it is easy, fast and one can decide for oneself what to read and when to read it. Whereas most participants had used the internet to obtain information, some participants did not have (or did no longer have) the urge to use the internet for health information because they believed it was too confrontational or led to unnecessary worry about their disease.

"It's fine by me, I can think of so many other things to search for and giving myself a hard time about. I live my life now and I don't want to think about it daily [Female, 40 years, RA]."

Also, many participants already felt that they knew everything they wanted to know. Some participants reported that they felt there is an overflow of information on the internet, which can make it hard to find relevant information, judge the reliability of information and interpret the information correctly. Other participants reported obtaining their information in alternative ways, such as through their doctor or from patient organization magazines. Information provision via a rheumatology IHCA from their own hospital provoked enthusiasm, since it could overcome the problem of information unreliability.

The information participants preferred the most could be classified into three categories. The first category is disease information, which contains topics such as the symptoms of the disease, the diagnosis, heredity and related symptoms, such as fatigue. Some patients mentioned that they want to be kept up to date on rheumatology research, to know about the latest results and developments.

"That is just keeping up with the newest developments within the field, as a patient. [Male, 55 years, Arthritis Psoriatica]."

Second, information about treatment was preferred, such as medication, therapies and protocols. The final category concerns information about how to cope with rheumatism, which involves topics such as dealing with the psychological and social consequences relating to family, friends and work, how to keep exercising, and tips and tricks to overcome the difficulties in daily life that rheumatism can cause.

2) Information about care and support

Participants were asked to what extend they used or were interested in a 'care guide': an overview of all the rheumatism care and support available in the region. Half of the participants reported knowing of, and using existing care guides. Participants thought that these tools gave structured and complete overviews of health care and support services, and that they were helpful in making informed choices concerning health professionals. The most important reason participants mentioned not to use a care guide was that they did not need any additional care, and if necessary current health professionals usually made recommendations. However, a care guide from a rheumatology IHCA from their own hospital would be appreciated by most participants; it was seen as a potentially good reference tool in healthcare and support.

"I used one (care guide, ed.) to find a physiotherapist in [small town] who was specialized in rheumatic diseases. Through this website I got the therapist I have now [Female, 53 years, RA]."

An effective care guide includes two kinds of information, according to the participants. The first type is aimed at medical care; the second type at support services and local resources. Regarding the medical care information, participants expect job descriptions and specializations of all health care facilities, including psychological and familial help. Each facility should show a complete overview of all its health care professionals. Also, information about accessibility, waiting periods, and hyperlinks to the web pages of each health professional is valued. A few participants would additionally like to read about experiences and opinions of other patients about particular professionals. Regarding the hospital participants wanted information about procedures, reciprocal expectations between the hospital and the patients, any changes in the rheumatism department and announcements of activities and meetings involving rheumatism. The preferred information on support services and local resources varied from household services to work reintegration authorities and health resorts for vacations. Participants also expressed a need for clear information about the options and financial help for home adjustments, support tools, health insurances and tips for disabled-friendly shopping, dining and entertainment in the region.

"It is not just the medical part that counts, but also the coping in daily life. Where can you find information? Which regulations are important for you? How are things covered financially? [Female, 57 years, Forestier's Disease]."

3) Communication with peers

One-third of the participants reported using online peer support groups or looking at support message boards occasionally. Participants identified advantages in online support groups since they can supply recognition, advice and support in coping with the disease. Furthermore, such groups

are anonymous, which reduces the reluctance to discuss personal topics.

"Larger issues you discuss with your doctor, but for me it is very nice to read about the little things and think 'oh, all those other people experience that too' [Female, 57 years, RA]."

Reasons for not using online support groups were that the information can be unreliable and some participants felt that people who are active in online support groups tend to complain a great deal or will only talk about their own problems. Also, some messages about the scope of the disease could be confrontational. Furthermore, some participants reported that they did not fit into the target group represented by the online support group.

"I searched a lot in the beginning, when I was just diagnosed with rheumatism, and then I stumbled upon a rheumatism peer support forum. That's when I thought that if this is where I'll end up, then I'm never looking again. I was really shocked by it [Female, 40 years, RA]."

"I know these (peer support groups, ed.), but they didn't appeal to me because there were mainly younger people posting on them, struggling with kids, getting married and jobs, but I already had all of that covered, so that wasn't an issue for me anymore [Female, 43 years, Ankvlosing Spondylitis]."

Some participants added that they perceived online communication as impersonal, that they didn't want to hear strangers' stories or advice, and that they didn't want to spend too much time reflecting on their disease; because they did not want to feel like being a patient all the time. However, because of the large amount and large diversity of pros and cons for peer support groups, most people found it difficult to give a clear opinion or preference about the desirability of such an application within a rheumatology IHCA.

"Personally, I don't want to be occupied with my disease too much. But on the other hand, I don't want to miss valuable advice [Female, 57 years, Forestier's Disease]."

Participants reported that there should be clear value for them in the online support groups: messages should be positive, and the exchange of tips and tricks should be the main function of the group. Other important pre-conditions were that there should be accurate control of posts as well as on privacy, and participants thought it was important to have a variety of topics and target groups on a forum.

To the question if a peer support group should be national or regional opinions split two ways. Half of the patients thought such a forum should be national, because they felt it could provide more information about how treatment and coping differs around the country. One participant even mentioned a world wide forum to learn more about current research and treatment development globally. Furthermore, patients saw more value in a national forum to be able to speak to new people; instead of to people they can also visit face-to-face.

"Yes, than I would use it more. See, I already have my contacts with rheumatic patients in the neighborhood [Male, 59 years, RA]."

Regional cultural differences were also mentioned; people in the region of Twente are known to be more introvert and down to earth than patients in southern or western regions of the Netherlands. A national forum could enrich the information flow, because more (different types of) patients can contribute. However, the other half of the patients reported to be more in favor of a regional forum. This might provide more recognition between patients from the same hospital, with the same doctors. Also, it would keep things small and orderly when not too many patients have access to the forum and can post messages. Furthermore, it would be easier to meet each other in person when desired. Strikingly, the regional characteristics were mentioned in this context as well; patients mentioned it would be nicer to talk to other patients who think and communicate alike.

4) Communication with the health professional

The majority of participants had never used e-mail to contact a doctor, either online available doctors or their own doctors. Almost all participants would never consider consulting an online doctor which they did not know, for it might be unreliable and it would feel as a betrayal to their own rheumatologist. However, there was a significant discrepancy between actual and preferred use of online contact with their own care provider in the hospital. Nearly all participants felt that this facility would be a valuable addition to the current care, since it is accessible, reliable and easy. Moreover, participants mentioned that e-mail allows them to take time to formulate a question and to carefully read or reread a doctor's answer. Participants would use e-mail mainly for minor, non-urgent questions, mostly instead of using the telephone to ask a question. Yet some patients mentioned it could stretch the time between two visits or it could possibly even save a visit to the hospital.

"Sometimes I just have a short question and it's not necessary to make an appointment. Something I just want to check. I don't have to make a telephone call for it either, there's no rush. Sending an e-mail would suffice [Male, 58 years, SLE]."

Despite their positive views, disadvantages were also mentioned: one disadvantage is the lack of immediacy in the communication, which inhibits both doctors and patients from directly asking a follow-up question for clarification. Also, patients would have to wait a while for a reply e-mail, while face-to-face or telephone contact is both direct and in real time.

Overall, participants thought that e-consultations could be a valuable extension of their current healthcare. The most important criteria for this tool are that the e-mail contact occurs with the rheumatology department of their own hospital and that it should not replace their regular contacts with their doctor. Moreover, participants expect a quick response of e-mails in a protected environment.

Some patients also mentioned to bundle abound questions to form a 'frequently asked questions'-tool (FAQ), or to create such a tool in advance, to avoid a lot of the same questions. Most patients reported they would use such a tool because the threshold would be very low. Furthermore, they reported that these tools often provide a lot of useful information. It could also contribute to the recognition that patients feel, because they are not the only one with those kind of questions.

"Those are things (FAQ's, ed.) I read a lot, and then I feel like 'oh, I am not the only one with these questions and they are already answered' [Female, 28 years, Fibromyalgia]."

Yet an important precondition is that the list of frequently asked questions does not become too large, which causes overkill and disrupts the orderly presentation of information.

Using a chat function to communicate with their care provider causes enthusiasm for almost half of the respondents. It would save patients the stress of visiting the hospital, including finding a parking space, sitting in the waiting room and absorb all the information of the doctor in one time. Still, it would be one step to far for a lot of patients. They are afraid the conversation would get too chaotic or they would not know how to use the tool properly. An important precondition would also be that the amount of offered chat sessions by the care providers could cover the demand by patients.

5) Participation by symptom monitoring

Half of the participants did not have experience with symptom monitoring. The other half had some experience in various ways, for example using a diary or during a treatment. Reasons mentioned for using a symptom monitoring tool were that it could give both the participant and the doctor a better insight into the disease, which could benefit communication and treatment. Also, it was considered to be good to be open-minded about new approaches and methods, and it can be fun to use the tool and see patterns emerge over time.

"You get a much better idea of what your bottlenecks are, and then you can explain it a lot better to the rheumatologist [Female, 40 years, Osteoarthritis]."

Some participants were not able to grasp the use and the extra value of regular monitoring. Often because they felt they did not have complaints that were severe enough to be worth monitoring or because their complaints had been stable for a longer period of time. Other reasons for not using symptom monitoring were that it could be confrontational, participants didn't want to spend too much time thinking about their disease and some patients feared it could be counterproductive if one becomes too focused on pain and symptoms.

"I just don't want to know. Ignorance is bliss; if I'm feeling good on a day, then I live it to the fullest. If I feel miserable the next day, then that's the way it is. I don't think about it too much [Female, 57 years, RA]."

Participants particularly appreciated the value of symptom monitoring when the data would not just be for their own knowledge, but when their doctor also receives the data and uses it to improve treatment. For example, a doctor could adjust treatment or medication according to reported complaints by patients, or a doctor could go deeper into conversation about monitored problems during consultation. An advanced way of getting the monitored information from the patient to the care provider would be via tele-monitoring. The IHCA could offer an application with which it is possible to get the patients' data directly to their own care providers. This way it could be a valuable addition to regular care.

"For me personally it would only work when it would benefit me, when I could improve something with it (symptom monitoring, ed.). But if I would only be scoring all my pains en symptoms, that would not do any good for me personally [Female, 56 years, Sjögren's Syndrome]."

Symptoms that participants would like to monitor were primarily inflamed and swollen joints. Furthermore, the monitoring of pain, overall health, and exercise is important to patients. Stress, fatigue, medication and nutrition were also mentioned. Furthermore, the participants thought it was important to see possible correlations between these various factors in graphs. For example the effect of exercising on perceived pain, or the effect of sleep on the amount of stiffness. Symptom monitoring would mainly be used semi-regularly in times of high disease activity, and before a consult. Participants also mentioned that it could be a very valuable tool for patients who were recently diagnosed, for exactly than it can give good insight in the variability and fluctuation of symptoms and pain.

6) Participation by exercise programs

Most participants did not have any experience with online exercise programs. They mentioned not having enough self-discipline to persist and they mostly preferred visiting the physiotherapist. Some participants regularly exercised by themselves, through daily activities such as walking or cycling, visiting the gym or using exercises from a self-help book or on a game computer. Still, these people

also mentioned having self-discipline problems in regulating their behavior.

"One time I got a booklet with exercises from the Dutch Arthritis Association. I started it, but on some point in time the motivation slipped away and I thought 'well, never mind then' [Female, 50 years, Osteoarthritis]."

Some participants reported that using an online exercise program might, in comparison to physiotherapy, help to maintain a sense of self-respect: doing things on your own. Furthermore, being able to exercise at home and get tips and support via the internet would be comfortable. Still, almost half of the participants did not see any value in an exercise program on a rheumatology IHCA. They did not think the tool could address the need for self-discipline. They were afraid of the accuracy and the safety of the exercises, and of doing them without a supervisor.

"It all depends on proper supervision. I can and I want to exercise, but if I do things the wrong way I get injured easily. When a healthy person does something incorrectly, he gets muscle aches, but if I do something incorrectly I can't walk for a week. To prevent this, I want a physiotherapist next to me. I want to keep on exercising, but in a healthy way [Female, 43 years, Ankvlosing Spondylitis]."

These barriers might be overcome by an online coach, someone who can watch the patient via a webcam, so that the coach can look along and give tips and advice. For some participants, this seemed like a good idea. Furthermore, patients would appreciate information and tips and tricks considering exercising. They would like to know which exercises are good, and which are not, or which could even be bad. They would also like advice on which exercises are good for what specific problems or for what specific parts of the body. Participants reported that this information would lower thresholds for them to start exercising in their own pace and convenience.

"I would indeed look up which exercises they recommended, and I can imagine that I would also actually use them [Female, 56 years, RA]."

7) Participation by access to medical health record

The most enthusiastically identified example of an online application by participants was access to their personal health record. Fifteen out of eighteen participants were positive about this; they would like to have access to their complete personal health record, including previous and current test and lab results, their treatment plans and an overview of all the upcoming appointments. The most important reason why they wanted this was to feel more involved with, and in control of, their disease and treatment.

"It would mean more involvement in myself. It concerns information about me, so I would like that very much (insight in personal health record, ed.) [Male, 59 years, RA]."

Also, it would give a good overview of the entire treatment, both back in time and in the future. Patients could see how their lab values and their disease activity have been changing over time, and they could see how their treatment is going to proceed and what they can expect from the hospital in the upcoming months.

"According to my treatment plan I have to give blood every 4 weeks and I have to get a consultation every 6 months. If I could see that in a schedule, I would never have to ask myself anymore 'How did this work again?' [Male, 59 years, Osteoarthritis]."

One reason for not desiring access to their personal health record would be that participants feel it is too difficult to understand all the information. Participants argue that it is the doctor's information and they would not know how to interpret it. Therefore, an important pre-condition is that the personal record should contain enough clear information and instructions to allow the patient to correctly interpret all the results and information.

"How is that score calculated and what is good or bad? I would not know, the nurse always scores everything and than says 'Well, you are doing fine' [Female, 39 years, RA]."

Furthermore, it is important that the records are safely secured. Patients want the information to be only accessible for themselves and for their care providers.

IV. DISCUSSION

To the best of our knowledge, this is the first study to identify a broad overview of use, needs, motives and preferences of rheumatism patients on a full spectrum of online support applications. Results reveal that the provision of an IHCA by ones' own hospital causes enthusiasm. With a hospital based IHCA barriers of online applications, such as information overflow or doubts about the reliability, could be overcome. Overall, participants were most interested in receiving information on both medical and support topics, online contact with their doctor and access to their personal health record. Patient support groups were less preferred as well as participation tools such as online symptom monitoring and online exercise programs. Furthermore, a significant discrepancy between current use and future preferences was seen in information about care and support, online communication with the doctor and access to a personal health record.

A. Information

Presently, participants used the internet predominantly to search for information. Previous research among rheumatism patients, as well as among those suffering from other diseases, has shown similar findings [25][26][27]. In this study the emphasis was on the kind of information and the

reasons for which patients searched the internet. We found that the participants were predominantly interested in disease information, treatment information and information on care, which was also reported by Gordon [28] and Hay [29]. Still, many participants reported searching for other information than the aforementioned subjects, which is not reflected in earlier studies. First, many participants emphasized information about coping: how to deal with psychological and social consequences relating to stress, family, friends and work, how to keep engaged in exercising, and tips and tricks to overcome the small difficulties in daily life that rheumatism can cause. Second, information on support services and local resources was valued, such as on household services and financial support for home adjustments. Overall, participants seem to want rheumatism information in a broader spectrum, while on the other hand many participants mentioned that they often experience an information overload. This is widely described in the literature, moreover, the available health information is often unreliable or biased [21][30]. Information provision by way of a hospital IHCA, could meet the preferences of patients by overcoming the problem of a doubtful information overload, while still offering a wide amount of information.

B. Communication

Participants were asked about online communication possibilities with both their doctors and their peer patients. The reported overall current use of communication tools by participants is limited. Much is written about the possible positive results peer support groups can give [31][32]. Still, actual usage of online support groups seems to be moderate [24][33][34][35][36]. The current study shows that most participants do not immediately reject the concept of online peer support groups, but they only want to participate under certain conditions. Participants would like to read positive messages and practical tips from other patients. Communication with health professionals shows a large discrepancy between current use and needs for the future. This is also shown by Van Lankveld in a study on current and expected use of online health applications by the chronically ill [27]. This discrepancy is largely due to lack of opportunity. Most participants have never communicated with their doctor online [33], because such applications were not available. Still, when offered, e-mail contact appears to be a popular facility [13][18][34][35][37]. Our study reveals that rheumatism patients thought it would be an accessible, reliable and easy way to improve their current care. However, patients do not want e-mail communication to replace consultations or other face-to-face contact. Moreover, practical implementation might be difficult as e-mail communication might be impeded by legal, budgetary and motivational barriers [38].

C. Participation

The current use of self-management or exercise programs is reported as moderate by patients. Many of our participants did not see the purpose of these applications, or they believed it would demand too great a time investment without clear benefits. Previous trial studies concerning self management and physical activity in rheumatism showed good results using computer-based technologies [15][31][39]. However, despite of promising results, the predicted use of suchlike tools on an IHCA is still moderate. Reported explanations for this are barriers in both self-discipline and accuracy, and safety of the exercises.

On symptom monitoring patients stated that their motivation to use the application would definitely increase if their doctor would use the information for treatment purposes. Therefore, the greatest promise of these tools is when linkage to the treatment can be realized. The perceived usefulness for themselves and for their treatment is a large motivator for use: this raises interesting questions about the possible future use of these applications. For example, tele-monitoring, in which the doctor applies a patient's self-reported data on monitoring and management during the consultation.

Finally, a participation tool with great potential is online access to medical health records. Previous studies have shown that this application is well received by patients [18][37][38][40] and the participants in this study also report enthusiasm. This application would give patients the sense of being involved in, and in charge of, their own disease and treatment. Motivations such as this are very important because they demonstrate the value an IHCA can have in involving patients in their care process.

D. Preferences

Results show that patients mentioned a lot of pre-conditions which should be fulfilled in order for them to use the various support tools, especially for the communication and participation tools (Table 2). Often, the reported motives for not using the support tools could be overcome when the patients' demands on possibilities, quality and care provider involvement for each online support tool are met. Therefore, it is very important to meet these pre-conditions when developing a rheumatology IHCA; this is what makes the application patient-centered. If we wish to overcome the current acceptance problems that many online applications face when it comes to actual use, patients' wishes should be complied with as much as possible and feasible. Still, it should be noticed that many patients have not yet had the opportunity to use most of the participation applications, so their preferences and pre-conditions are not based on experience, but on expected usefulness [41].

E. Study limitations

There are limitations to this study. This qualitative study may not be representative for all patients. The participants were volunteers who, being more actively involved in research than usual patients, may not represent typical patients. Furthermore, the participants had mostly suffered rheumatism for a longer time. This can influence their needs and preferences; they are in a later stage of their illness

journey than recently diagnosed patients. In a quantitative follow-up study these limitations will have to be averted.

V. CONCLUSION

Patients see great value in an IHCA provided by their own hospital, since it could increase reliability of the provided information , and would give them the confidence to use the application. The current study shows a significant discrepancy between current use and future preferences rheumatism patients have regarding online communication with their doctor, online symptom monitoring and access to their medical health record. Furthermore, our results provide an overview of important preferences and pre-conditions that patients have for each support tool in order to improve intention to use the application. Overall, patients prefer a rheumatology IHCA that contains both communication and participation tools, which are linked to the hospital, and information about disease, care and practical support.

ACKNOWLEDGMENTS

We would kindly like to thank all the patients that participated in this study for their time and effort. This work was supported by an unrestricted educational grant from Wyeth Pharmaceuticals, part of Pfizer. This funding source had no involvement in data collection, analysis, or the preparation of this manuscript.

REFERENCES

[1] Van der Vaart, R., Drossaert, C.H.C., Taal, E., Van de Laar, M.A.F.J. (2010). Experiences and preferences of patients with rheumatic diseases regarding an interactive health communication application. *eTELEMED'10 Second International Conference on eHealth, Telemedicine, and Social Medicine,* 64-71 doi: 10.1109/eTELEMED.2010.16.

[2] Demiris, G., Afrin, L.B., Speedie, S., Courtney, K.L., Sondhi, M., Vimarlund, V., Lovis, C., Goossen, W. and Lynch, C. (2008). Patient-centered applications: use of information technology to promote disease management and wellness. A white paper by the AMIA knowledge in motion working group. *Journal of the American Medical Informatics Association, 15,* 8-13.

[3] Keselman, A., Logan, R., Smith, C.A., Leroy, G. and Zeng-Treitler, Q. (2008). Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association, 15,* 473-483.

[4] Powell, J.A., Darvell, M., Gray, J.A.M. (2003). The doctor, the patient and the world-wide web: how the internet is changing healthcare. *Journal of the Royal Society of Medicine, 96.*

[5] Ball, M.J. and Lillis, J. (2001). E-health: transforming the physician/patient relationship. *International Journal of Medical Informatics, 61,* 1-10.

[6] Gustafson, D.H., Hawkins, R., Boberg, E., Pingree, S., Serlin, R.E., Graziano, F. and Chan, C.L. (1999). Impact of a patient-centered, computer-based health information/support system background: Consumer health information systems potentially improve a patient's quality of life and activate patient self-care. *American Journal of Preventive Medicine, 16*(1), 1-9.

[7] Lapsley, P. and Groves, T. (2004). The patient's journey: travelling through life with a chronic illness, authors. *British Medical Journal, 329,* 582–583.

[8] Kerr, C., Murray, E., Stevenson, F., Gore, C. and Nazareth, I. (2006). Internet interventions for long-term conditions: patient and caregiver quality criteria. *Journal of Medical Internet Research, 8*(3).

[9] Winkelman, W.J. and Choo, C.W. (2003). Provider-sponsored virtual communities for chronic patients: improving health outcomes through organizational patient-centered knowledge management. *Health Expectations, 6,* 352-358.

[10] Nguyen, H.Q., Carrieri-Kohlman, V., Rankin, S.H., Slaughter, R. and Stulbarg, M.S. (2004). Internet-based patient education and support interventions: a review of evaluation studies and directions for future research. *Computers in Biology and Medicine, 34,* 95–112.

[11] Murray, E., Burns, J., See Tai, S., Lai, R. and Nazareth, I. (2009). Interactive health communication applications for people with chronic disease (Review). *Cochrane Database of Systematic Reviews, 1.*

[12] Garcia-Lizana, F., Sarria-Santamera, A. (2007). New technologies for chronic disease management and control: a systematic review. *Journal of Telemedicine and Telecare, 13,* 62–68.

[13] Solomon, M.R. (2008). Information technology to support self-management in chronic care: a systematic review. *Disease Management Health Outcomes, 16*(6), 391-401.

[14] Ansani, N.T., Fedutes-Henderson, B.A., Weber, R.J., Smith, R., Dean, J., Vogt, M., Gold, K., Kwoh, C.K., Osial, T. and Starz, T.W. (2006). The Drug information center arthritis project: providing patients with interactive and reliable arthritis internet education. *Drug Information Journal, 40*(1), 39-49.

[15] Stinson, J.N., Stevens, B.J., Feldman, B.M., Streiner, D., McGrath, P.J., Dupuis, A., Gill, N. and Pertroz, G.C. (2008). Construct validity of a multidimensional electronic pain diary for adolescents with arthritis. *Pain, 136,* 281-292.

[16] Van den Berg, M.H., Ronday, H.K., Peeters, A.J., Le Cessie, S., Van der Giesen, F.J., Breedveld, F.C. and Vliet Vlieland, T.P.M. (2006). Using internet technology to deliver a home-based physical activity intervention for patients with rheumatoid arthritis: a randomized controlled trial. *Arthritis & Rheumatism, 55,* 935-945.

[17] Lorig, K.R., Ritter, P.L., Laurent, D.D. and Pland, K. (2008). The internet-based arthritis self-management program: a one-year randomized trial for patients with arthritis or fibromyalgia. *Arthritis & Rheumatism, 59,* 1009-1017.

[18] Eysenbach, G., Fartasch, D. and Diepgen, T.L. (1999). Evidence-based patient education on the web: methods and studies for determining consumers' needs. *Journal of Medical Internet Research, Supplement 1.*

[19] Silvestre, A., Sue, V.M. and Allen, J.Y. (2009). If you build it, will they come? The Kaiser permanente model of online health care. *Health Affairs, 28*(2), 334-344.

[20] Or, C.K.L., Karsh, B. (2009) A systematic review of patients acceptance of consumer health information technology. *Journal of the American Medical Informatics Association, 16,* 550–60.

[21] Wilson, E.V. and Lankton, N.K. (2004). Modeling patients' acceptance of provider-delivered e-health. *Journal of the American Medical Informatics Association, 11*(4), 241-248.

[22] Klein, R. (2007) An empirical examination of patient-physician portal acceptance. *European Jouranl of Information Systems, 16,* 751–60.

[23] Guest, G., Bunce, A. and Johnson, L. (2006). How many interviews are enough?: an experiment with data saturation and variability. *Field Methods,18*(1), 59-82.

[24] Patton, M.Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA:Sage.

[25] Eysenbach, G. and Köhler, C. (2003). What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analyses of search engine queries on the internet. *AMIA 2003 Symposium Proceedings, page 225.*

[26] Richter, J.G., Becker, A., Specker, C., Monser, R. and Schneider, M. (2003). Disease-oriented internet use in outpatients with inflammatory rheumatic diseases. *Zeitung fur Rheumatologie, 63,* 216-222.

[27] Van Lankveld, W.G.J.M., Derks, A.M. and van den Hoogen, F.H.J. (2005). Disease related use of the internet in chronically ill adults: current and expected use. *Annals of the Rheumatic Diseases, 65,* 121-123.

[28] Gordon, M.M. and Capell, H.A. (2002). The use of the internet as a resource for health information among patients attending a rheumatology clinic. *Rheumatology, 41,* 1402–1405.

[29] Hay, C.M., Cadigan, R.J., Khanna, D., Strathmann, C., Lieber, E., Altman, R., McMahon, M., Kokhab, M. and Furst, D.E. (2008). Prepared patients: internet information seeking by new rheumatology patients. *Arthritis Care & Research, 59*(4), 575-582.

[30] Glenton, C., Paulsen, E.J. and Oxman, A.D. (2005). Portals to wonderland: health portals lead to confusing information about the effects of health care. *Medical Informatics and Decision Making, 5*(7).

[31] Shigaki, C.L., Smarr, K.L., Gong, Y., Donovan-Hanson, K. and Siva, C. (2008). Social interactions in an online self-management program for rheumatoid arthritis. *Chronic Illness, 4,* 239–246.

[32] Eysenbach, G., Powell, J., Englesakis, M., Rizo, C. and Stern, A. (2004). Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *British Medical Journal, 328*.

[33] Anderson, J.G. (2004) Consumers of e-health: patterns of use and barriers. *Social Science Computer Review, 22,* 242.

[34] Siva, C., Smarr, K.L., Donovan Hanson, K., Parikh, M., Lawlor, K. and Ge, B. (2008). Internet use and e-mail communications between patients and providers: a survey of rheumatology outpatients. *Journal of Clinical Rheumatology, 14,* 318-323.

[35] Atkinson, N.L., Saperstein, S.L. and Pleis, J. (2009). Using the internet for health-related activities: findings from a national probability sample. *Journal of Medical Internet Research, 11*(1).

[36] Van Uden-Kraan, C.F., Drossaert, C.H.C., Taal, E., Lebrun, C.E.I., Drossaers-Bakker, K.W., Smit, W.M., Seydel, E.R. and Van de Laar, M.A.F.J. (2008). Coping with somatic illnesses in online support groups: do the feared disadvantages actually occur? *Computers in human behavior, 24*, 309-324.

[37] Weingart, S.N., Rind, D., Tofias, Z. and Sands, D.Z. (2006). Who uses the patient internet portal? The PatientSite experience. *Journal of the American Medical Informatics Association, 13*(1), 91-95.

[38] Hassol, A., Walker, J.M., Kidder, D., Rokita, K., Young, D., Pierdon, S., Deitz, D., Kuck, S. and Ortiz, E. (2004). Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging. *Journal of the American Medical Informatics Association, 11* (6).

[39] Lorig, K.R., Ritter, P.L., Laurent, D.D. and Plant, K. (2006). Internet-based chronic disease self-management: a randomized trial. *Medical Care, 44,* 964–971.

[40] Dorr, D., Bonner, L.M., Cohen, A.N., Shoai, R.S., Perrin, R., Chaney, E. and Young, A.S. (2007). Informatics systems to promote improved care for chronic illness: a literature review. *Journal of the American Medical Informatics Association, 14*, 156-163.

[41] Flynn, D., Gregory, P., Makki, H., Gabbay, M. (2009). Expectations and experiences of eHealth in primary care: a qualitative practice-based investigation. *International Journal of Medical Informatics, 78,* 588–60.

Learning Contexts as Ecologies of Resources: A Unifying Approach to the Interdisciplinary Development of Technology Rich Learning Activities

Rosemary Luckin
The London Knowledge Lab
The Institute of Education
London, England. WC1N 3QS
r.luckin@ioe.ac.uk

Abstract - This paper addresses the problem of how to develop a conceptualization of context that can support the development of technology-rich learning activities. The term *technology-rich* encompasses mobile, hybrid and on-line learning approaches and the work reported here is intended to bridge these different approaches. In this paper we suggest that a learner-specific definition of context can ground research across mobile, hybrid and on-line learning. We discuss a definition of context that is theoretically grounded in the socio-cultural approach to learning and that has been used to develop the Ecology of Resources model. This model is an abstraction that can be shared between social and technical researchers and practitioners to support analysis and to generate technology design. An example that demonstrates the way that the Ecology of Resources model is empirically as well as theoretically grounded is presented. This example is used to support the proposal that the Ecology of Resources model can be used as a design tool to sensitize designers to the importance of each learner's context.

Keywords - context, zone of collaboration, ecology of resources model, distributed scaffolding.

I. INTRODUCTION

The continuing increase in the range of pervasive, interconnected, and embedded technologies in our environment allows people to digitally link their experiences across, between and with multiple locations, multiple people and a range of subject matter. These technical developments have the potential to support the better integration of learners with their social, physical and digital worlds. Or in other words these developments have the potential to enable us to take batter account of each learner's context. The aim of this paper is to discuss the concept of context and to evaluate a context-based model of learning that is intended to support the development of technology-rich learning activities. The model is called the Ecology of Resources and it offers a potentially unifying concept for the sub-fields of learning with technology, both technical and sociological. In particular, the Ecology of Resources model aims to engender the development of activities that use technology to overcome the traditional physical and temporal constraints that are part of many learning environments and that are at the heart of approaches within the sub-fields of mobile, hybrid and on-line learning.

In this paper we extend [1] and discuss the Ecology of Resources model to consider some of its theoretical grounding. We then present an empirical example of the model in use. This example is drawn from the Homework research project: a project that explored the use of multiple technologies to support young learners (aged 5-7 years) with numeracy both inside and outside school. Such a situation is particularly compatible with the aim of the Ecology of Resources to support the development of activities that use technology to overcome the traditional physical and temporal constraints. This type of empirical evaluation adds both to our understanding of learners' interactions with technology, and to the continuing development of the Ecology of Resources model and design approach.

II. CONTEXT AND LEARNING

There is nothing new about the suggestion that one should explore the educational context in which learning takes place in order to understand more about learning. Work such as that completed by [2] suggests that the organization of learning resources, including the computer, influences the manner in which these resources are used. Similarly [3], when evaluating Integrated Learning Systems, concluded that the impact of technology upon learning was heavily dependent on the specifics of the educational environment into which the technology was introduced. This type of work is useful in confirming the importance of looking at the wider environment, but is limited by a focus that is mainly on specific environmental locations, such as school classrooms. To make the most of the possibilities afforded by new technologies *a clearer, learning-specific definition of context* is now required, to support the development of technology-rich learning activities. Activities that take advantage of the growing range of

technological artefacts that can support interaction across multiple physical and virtual spaces, multiple knowledge domains, multiple time periods and with multiple collaborators. The provision of such a definition is not an easy task. Nardi, [4] states the problem clearly: 'How can we confront the blooming, buzzing confusion that is "context" and still produce generalizable research results?'

In [5], we confront this "confusion" by looking at a range of ways in which context is talked about within literature drawn from multiple disciplines to identify common themes of concern that transcend disciplinary boundaries. This encompasses work drawn from geography and architecture, anthropology and psychology and from education and computer science. We conclude that:

> "Context matters to learning; it is complex and local to a learner. It defines a person's subjective and objective experience of the world in a spatially and historically contingent manner. Context is dynamic and associated with connections between people, things, locations and events in a narrative that is driven by people's intentionality and motivations. Technology can help to make these connections in an operational sense. People can help to make these connections have meaning for a learner.
>
> A learner is not exposed to multiple contexts, but rather has a single context that is their lived experience of the world; a 'phenomenological gestalt' [6] that reflects their interactions with multiple people, artefacts and environments. The partial descriptions of the world that are offered to a learner through these resources act as the hooks for interactions in which action and meaning are built. In this sense, meaning is distributed amongst these resources. However, it is the manner in which the learner at the centre of their context internalizes their interactions that is the core activity of importance. These interactions are not predictable but are created by the people who interact, each of whom will have intentions about how these interactions should be. " [5]

This definition portrays context as something that is centred around an individual. This results in a conceptualization of context with a time-scale that is an individual's life and boundaries that are those of the individual's interactions. We suggest that this definition of context can be used to ground the development of technology rich learning activities that do not differentiate between the various flavours of learning that inhabit the growing rhetoric of descriptors that include mobile, hybrid, virtual, on-line, and e-learning. All are concerned with learning and all can be supported by such a learner centric definition of context. But what theory of learning is consistent with this view of context and capable of being used to develop a context-based model of learning that can be made operational and that can form the foundation for a design framework?

There are several theoretical viewpoints that specifically relate to learning and context; for instance, work from the socio-cultural tradition, such as that of Vygotsky, activity theory, Michael Cole's cultural psychology, Hutchins' distributed cognition and the situated and communities of practice approaches. The discussion of context above favours an intentional role for people, a learner-centredness that defines context as an interactional concept, combining a learner's active experience of their physical reality with their mediated experiences through human-made artefacts.

The process of internalization through which an individual's distributed meaning-making interactions lead that individual's development is key to this enterprise. This narrows down the compatibility of the potential learning theories to those from a socio-cultural stance. The relational attributes of activity systems certainly make them appealing for this purpose. However, it is my intention to focus upon the learner at the centre of their interactions we therefore consider in more depth the socio-cultural approach of Vygotsky [7] [8]

A. Vygotsky, Learning and Context

The socio-cultural approach of Vygotsky is developmental and describes an individual's mental development as an interaction between that individual and their socio-cultural environment. The nature of these interactions influences the nature of their resultant mental processes; the interpsychological becomes intrapsychological via the process of internalization [7].

This approach offers compatibility with a context-based model of learning and is further focused in the Zone of Proximal Development (ZPD), which can be described as the crystallization of the internalization process. When Vygotsky [7] introduced the ZPD, he proposed that a child's ability to solve standardized problems unassisted was not the whole story of their development, but rather it simply reflected the completed part of their development. The ZPD was defined as:

> "The discrepancy between a child's actual mental age and the level he reaches in solving problems with assistance indicates the zone of his proximal Development; … Experience has shown that the child with the larger zone of proximal development will do much better in school. " [7]

Vygotsky suggested that when instruction is aimed at what the child can achieve when aided by a more able partner then it can play a major role in the development of that child's higher mental processes. The purpose of the ZPD is to focus the dialogue between the more able partner and the child, so that the learner can reflect upon this dialogue and reformulate it into their own thought [10].

The ZPD has an important *process* element as well as a conceptual one. This process element can be seen more clearly in Vygotsky [8], in which the ZPD is described as something that must be created through instructional interactions and that can only operate when the child is interacting with other people in the environment.

The ZPD could be thought of as the basis for a context of productive interactivity. However, the ZPD is underspecified; it does not, for example, specify the manner in which the 'actual developmental level and the zone of proximal development' [8] are to be identified. This need for clarification has been recognized by many researchers, such as [11 and 12]. Work such as the scaffolding metaphor of [13] and the constructs proposed by [14] have provided useful clarifications about how a ZPD might be created and have informed the interpretation of the ZPD that this paper discusses and that is at the heart of the Ecology of Resources.

My interpretation explores the relationship between the identification of a learner's collaborative capability and the specification of the assistance that needs to be offered to the learner in order for them to succeed at a particular task. We refer to this interpretation as the Zone of Collaboration, which uses two additional constructs called:

The Zone of Available Assistance (ZAA);
and
The Zone of Proximal Adjustment (ZPA).

The ZAA describes the variety of resources within a learner's world that could provide different qualities and quantities of assistance and that may be available to the learner at a particular point in time. The ZPA represents a sub-set of the resources from the ZAA that are appropriate for a learner's needs. Figure 1 represents these concepts graphically.

B. Scaffolding

As can be seen from the discussion so far, the ZPD requires assistance for the learner from a more able other. The nature of this assistance was, however, left underspecified in Vygotsky's writing. Seminal work done by David Wood [9], in which he coined the term 'Scaffolding' to describe tutorial assistance, is particularly relevant here. Effective scaffolding is presented as something more than the provision of hints and graded help. It involves simplification of the learner's role and interactions in which learners and their more able partners work together to achieve success, but the contributions from each vary according to the child's level of ability [15].

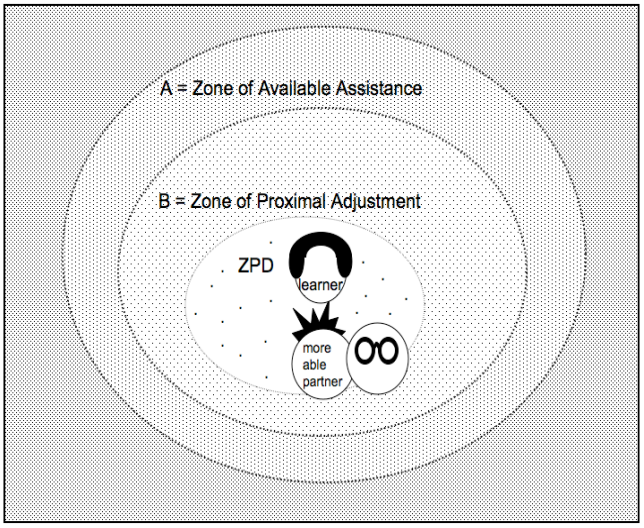


Figure 1 The Zone of Collaboration [5]

The scaffolding approach has been used to develop a variety of educational software, such as that of Wood, Shadbolt, Reichgelt, Wood & Paskiewitz [16], and has been used, adapted and extended to guide the development of a variety of technology enhanced learning applications (see for example, [17, 18, 19]). A further dimension to the use of technology for scaffolding can be seen in the use of scaffolding to support the development of higher order thinking skills, such as metacognition [20] and help-seeking skill development [21].

The potential offered by connected technology and the need to focus on context as discussed in the introduction to this paper require that consideration is given to resources that are beyond those offered by a single interactive learning environment, which is where most of the research attention has been focused to date. Some consideration has however started to be given to the possibilities afforded by scaffolding in these much more complex environments. [22], for example, use the term 'distributed scaffolding' and explore this through classroom-based science learning. Key findings from their work include identification of the increased complexity that occurs when scaffolding is distributed and the potential for distributed scaffolding to offer learners more opportunities to notice scaffolding opportunities. Tabak [23] also explores complex settings and distributed scaffolding and also identifies this positive possibility of increased opportunities for scaffolding. Her vision for distributed scaffolding is that learners can take advantage of different types of support provided by different means in an integrated manner, in order to solve complex problems. This notion of distributed scaffolding is increasingly important when a learner's broader context beyond a single learning environment is considered.

We therefore add to the definition of context that:

> "it is the role of the more able participants to scaffold a learner's construction of a narrative that makes sense of the meanings distributed amongst the resources with which they interact. Through this scaffolding the learner at the centre of their context internalizes their interactions and develops increased independent capability and self-awareness."
>
> [5]

III. THE ECOLOGY OF RESOURCES MODEL OF CONTEXT

The Ecology of Resources model builds upon this definition and develops the ZAA and ZPA concepts into a characterization of a learner and the interactions that form that learner's context. An earlier version of the model is discussed in [24] and its full detail can be found in [5]. Here we describe it briefly in order to ground the presentation of an empirical example and to support the suggestion that it might act as a useful mediating artefact to integrate work across various subfields such as mobile and hybrid learning (see Figure 2 for an illustration of the Ecology of Resources model).

The resources that comprise a learner's ZAA embrace a wide range of types, including people, technologies, buildings, books and knowledge. It is useful to consider the different types or categories of resource that might be available in order to help us identify them and the relationship they bear to the learner and to each other. One of the resource categories that the learner needs to interact with comprises the 'stuff that is to be learnt': the knowledge and skills that are the subject of their learning. A second category of resource is that described as 'Tools and People'. This category includes books, pens and paper, technology and other people who know more about the knowledge or skill to be learnt than the learner does. The last category of resource is that represented by the 'Environment' descriptor. This category includes the location and surrounding environment with which the learner interacts: for example, a school classroom, a park, or a place of work. In many instances, there is an existing relationship between the resources within these three categories: Knowledge and Skills, Tools and People and Environment. For example, the book resources appropriate for learning French are located in the Language Learning section of the library and formal lessons probably take place in a particular location in school. Hence, in Figure 2 the categories of resource surrounding the learner, and with which they interact, are joined together. In order to support learning, the relationships between the different types of resource with which the learner interacts need to be identified and understood. They may need to be made explicit to the learner in order to build coherence into the learning interactions. For example, if we wish to teach French conversation to an evening class, which involves how to order a meal, we may choose to provide a menu and to organize the room like a restaurant. We will also need to ensure that the language concepts we introduce are relevant to meal ordering.

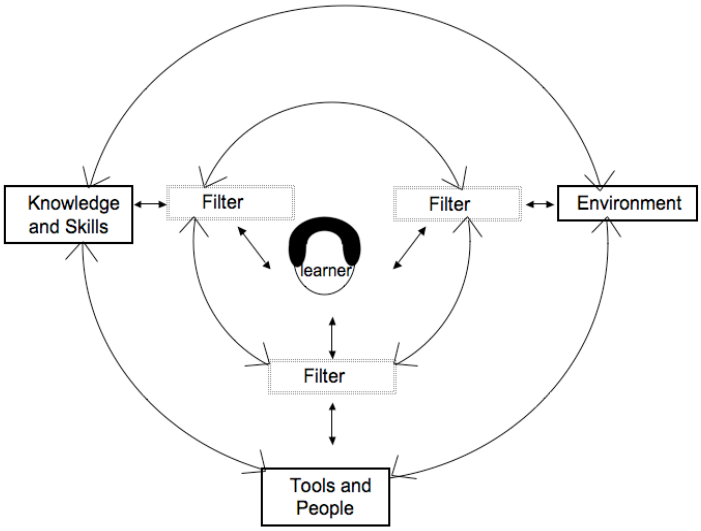


Figure 2 The Ecology of Resources Model [5]

A. *The Ecology of Resources Filter Elements*

This language-learning example highlights another factor that needs to be taken into consideration. We mentioned that we might organize the room in a particular fashion. This is an example of the way in which a learner's interactions with the available resources are often filtered by the actions of others — in this case, me as the teacher — rather than experienced directly and unimpeded by the learner. For example, the subject matter to be learnt is usually filtered through some kind of organization, such as a curriculum, that has been the subject of a process of validation by other members of the learner's society. This resource filter is stronger for subjects such as mathematics and other formal educational disciplines than for more grounded skills such as farming. However, even with skills-based subjects there is, to some extent at least, still some formalization of what is recognised as the accepted view about the nature and components of the skills that need to be mastered. The tools and people that may be available to the learner are also organized or filtered in some way. For example, a teacher taking a French conversation evening class is only available during that class, or perhaps at some other times via email. Classroom technologies are not always available to learners whenever they want: there are school rules and protocols that restrict the learner's access to resources. Finally, and again as reflected in the French conversation learning example, a learner's access to the Environment is mediated by that Environment's Organization. This resource filter is more obvious in formal settings such as schools, where timetables and regulations have a strong influence on the ways in which learners interact with their environment. In the same way that there may already exist a relationship between the different

resource elements in the outer circle of Figure 2, there may also exist a relationship between the filter elements. The coherence of the learner's experience can be enhanced through careful consideration of the existing relationships between the filter elements and between the individual resource elements and their associated filter. All of the elements in any Ecology of Resources bring with them a history that defines them, as well as the part they play in the wider cultural and political system. Likewise, the individual at the centre of the Ecology of Resources has their own history of experience that impacts upon their interactions with each of the elements in the Ecology.

IV. THE ECOLOGY OF RESOURCES MODEL IN USE

The Ecology of Resources model offers a way to talk about learners holistically – to sensitize us to the range of interactions that constitute their contexts, to frame the participatory design process; to explore data to understand more about learners' contexts; to identify the assistance that could be available and the way that learners' interactions with it might be filtered and supported and to identify situations where scaffolding might be used. The Ecology of resources approach has been used in a variety of projects that include science learning in school, informal and formal learning in the developing world and home education in the UK.

The example we draw upon here is that of the Homework project through which a system called HOMEWORK was developed. This was an interactive mathematics education system for children aged 5–7 years. The system used a combination of interactive whiteboard and Tablet PC technology, plus some bespoke software, consisting of lesson planning, control and home use components. The system contained a rich set of multimedia and associated interactive numeracy resources drawn from the popular television series called the Number Crew. Teachers used the software to link resources into lesson plans. In the classroom, the interactive whiteboard was used for whole class activities and each child also had their own Tablet PC for individual and small group activities. The teacher could control the classroom activity from their own Tablet PC and could allocate new activities or send messages to individuals or groups of children in real time. When planning each lesson the teacher could also decide upon homework activities and allocate them to individual children's Tablets as appropriate. After school, the children took their Tablet PC home with them and used it at home or elsewhere; individually or with parents. At home, in addition to homework activity set by the teacher, the Tablet provided access to the resources the learner had used in class that day, the resources that they had used in previous sessions (irrespective of whether the child was actually in school or not) and information for parents about the learning objectives to which these activities related. There were also links to other relevant fun activities and a messaging system to support parent and teacher communication. The HOMEWORK system was developed incrementally and interactively with learners, teachers and parents. Each iteration gave the design team a clearer understanding of the interactions that made up the learning contexts of the children who the system was developed to support. It is described in some detail in [24]. Here, we use the example of the system that was the product of this development and its empirical evaluation to demonstrate its use of mobile, classroom and e-learning technologies and the manner in which the Ecology of Resources model of context can be used to model learner interactions and design technology use.

Evaluations of the HOMEWORK system were conducted in different schools and classes throughout the system's development. The evaluation on which we draw in this paper was conducted with a class of 32 children aged 5–7 years. The research was exploratory and was concerned with understanding the nature of the learning interactions that learners, teachers and parents were able to engage in supported by the HOMEWORK system. There was no intention to set up a comparative control group trial.

Multiple data sources were collected which included: logs maintained by the system; diaries maintained by parents; interviews with parents; and questionnaires completed by parents. In this example we focus upon the data that illustrate how the Tablets were used by children and their families outside the classroom. It is these data that can offer valuable information about the child's wider learning experience across multiple locations, tools and with a variety of other people.

The HOMEWORK system was used for three, hour-long mathematics lessons per week in school. The log data indicate that the Tablets were used at home, on average, slightly less than once a day for the equivalent of 25 minutes a day. However, there was great variability in both session length and in the total time children spent using the Tablets during the research period. The diaries maintained by parents indicate that the most common time for the Tablet to be used was weekday evenings (after 5.30 pm) and during the daytime at weekends. These diaries also reveal that the Tablet was most often used at a table located in a communal space such as a lounge, and that mum was the person who most frequently helped children with their Tablet activities.

There was also evidence of learning gains during the time the HOMEWORK system was in use. These can be seen in the changes in children's scores in a pre-test and post-test set by the teacher. The mean scores for the youngest children (5–6 years of age) increased by 17 per cent between the start of the study (T1) and the end of the study (T2); and for the older children (6–7 years of age) by 26 per cent, as illustrated in Figure 3.

In addition to these test scores, parents' comments in the diaries they maintained and during interviews also suggest that children's learning may have benefited during the

Homework project studies. For example, a parent commented about her daughter Jane:

> "Jane wanted to go through the videos of the number crew. Jane has definitely found having the PC a more interesting way of doing maths, as with most things, seeing and having an active part to play means more than just listening. Jane said that she had learnt a lot this weekend and showed me ½s and = and 10s and 100s counting on and backwards and you could tell how pleased she was with herself. Jane cleaned the PC and is taking great care of it."

In the interviews with the class teacher there are also reports that reflect the teacher's belief that children are benefiting from their use of the HOMEWORK system. The findings from the interviews, both with parents and teachers, provide confirmation of the pre-test/post-test data.

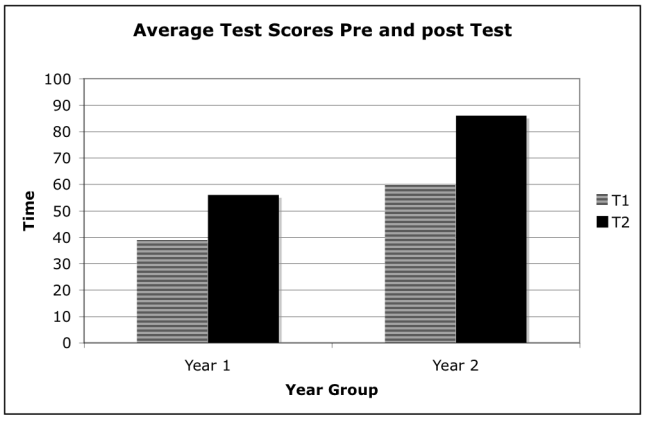


Figure 3 Average test scores pre and post system use

A. Family Case Study

In addition to this data about learning gains and how, when and where the tablets where used, the project team also wanted to explore more about the nature of the way in which the system may have supported learning, and the range of interactions that the HOMEWORK system was able to support. The multiple data sources were therefore pulled together to produce narrative family case studies. We discuss an excerpt from one such case study here to illustrate one family's use of the system in more depth. This narrative is about a learner — Robert, and his family — and the way that they used the HOMEWORK system through the Homework Tablet. The narrative is constructed from the data logged by the Homework Tablets and the entries made in the diary kept by the family. The timings for the length spent on activities are rounded down to the nearest minute. The '…' symbol indicates where there is a day's entry in the full narrative that is not part of this extract.

1) Robert's Story (an excerpt)

Robert is 5 years old and lives at home with his parents and two brothers.

It's Wednesday and Robert has used his Tablet PC in school this morning, from 11.20am, working on an activity about estimating a number of objects. He watched a video to begin with and then worked through activity number 7. He completed activity number 7 at level 1 for 5 minutes, skill number 7 at level 1 for 3 minutes, activity number 7 at level 2 for 2 minutes, skill number 7 at level 2 for 3 minutes and activity number 7 at level 3 for a little over 2 minutes. He finished at 11.45am.

Robert takes his computer home this evening and uses it at 6.10pm. He works in the lounge, sitting on the floor with his Tablet PC on the coffee table, with his grandmother and brother. He works on the 'Number Themes' digital camera homework for about 15 minutes (6.10pm – 6.25pm). This asked him to look for numbers in his home and to take photographs using the camera integrated into the Tablet PC. He takes 2 photos, which do not come out very well. Mum says there was not enough light to take photos (in diary). He then has a look at the other homework activity called 'Number Bags' at about 6.35pm.

…

Friday - Robert's teacher uses the HOMEWORK system for the maths lesson today and starts the session with the whole class together sitting on the floor in front of the interactive whiteboard singing the Number Crew song, watching a video and completing a numeracy activity. Each child then returns to their seat and uses their individual Tablet PC. Robert uses his Tablet PC this morning at 11.45am for about 20 minutes. The teacher has set some homework and Robert looks at this while the teacher explains it to the class. The homework activities for this week are called 'Numbers are everywhere', which involves using the Tablet PC camera to take pictures out of school of numbers up to 100 around the house and then complete a worksheet; an interactive activity called 'Ten Down'; and a video called 'Storm and Seasickness 1'. Robert has a practice with the camera, for about 10 minutes, and takes and looks at some pictures.

Robert takes his Tablet PC home for the weekend today. He uses it when he arrives home from school at 3.45pm. He works at the kitchen table with his brother. He looks at the fun activities and the homework. He then spends 25 minutes doing the 'Ten Down' homework activity (see Figure 4), from about 3.50pm – 4.15pm and briefly looks at the 'Numbers are everywhere' camera homework sheet. He turns on his PC again at 5.20pm and once again

works at the kitchen table, this time with Mum. He looks at the 'Numbers are everywhere' homework again, has a 2-minute go with the camera, during which he records a video and then has another little play with the 'Ten Down' activity for 10 minutes. This involves a lot of exploration as Robert looks at a number of different activities for a few minutes each. He keeps coming back to the Tablet PC: he plays for about 20 minutes at 5.35pm, then 10 minutes at 6.10pm, then again for about 10 minutes at 6.50pm. Mum says he really enjoyed the 'Ten Down' activity and grasped it quickly. The day's session finally ends at 7pm.

Saturday Robert uses his PC again on Saturday morning, turning it on briefly at 8.40am when he returns to the 'Ten Down' exercise for 5 minutes, working in the lounge. At 10.30 am he works at the kitchen table with his Nanny and wants to repeat the 'Ten Down' activity as he had enjoyed it. He also explores a little more and looks at the fun activities and the activities that he did at school yesterday. He has another look at the 'Numbers are everywhere' worksheet, and takes a picture with the Tablet PC camera.

Figure 4 Ten Down activity

Robert's story offers an unusual and valuable insight into technology use out of school. It illustrates that he made use of the flexibility offered by the technology and used his Homework Tablet in a variety of locations, and at different times throughout the day. He could choose when and where to work on his numeracy within the constraints negotiated with his family. Sometimes he worked on an activity for a minute or two and on other occasions for longer. Robert worked on the homework activities set by the teacher, but did more besides and was able to choose what he wanted to work on, could show it to his other family members and, in so doing, behaved independently. He used the Tablet PC review activities completed in the past, both at school and at home: these activities might be whole class, small group or individually based and might use the fixed whiteboard technology or the mobile Tablet PC technology.

The nature of the technology is not the differentiating factor of Robert's learning experience. The homework system, with its combination of technologies, linked each learner's experiences at school with their experiences outside school, and helped provide conceptual coherence, so that the knowledge learnt at school was made relevant for home, too, and not seen as something that was only for formal school education. The empirical analysis also highlighted the fact that the experience that a learner has, when using a particular piece of content, is part of its personalization for that learner. So whilst two learners may both watch the same video clip or complete the same worksheet in their homes, their experience of this will be different owing to the interactions that surround their experience, such as the conversations they have with their sibling about the video clip or the comment that Mum makes whilst they are completing the worksheet. There can be no assumptions made about content that works well in school working equally well, or in the same way, when this same content is used outside school. The content needs to be adapted if the learning interactions we want learners to engage in are to be integrated both inside and outside school.

One of the important possibilities afforded by new technology is that it can also link together the people who are acting as MAPs in the different locations that comprise a learner's context. In the Homework project example this would mean linking parents and family members with the teacher and assistant, for instance. In fact this was achieved through various mechanisms: for example, there was a messaging service available with the tablet PC through which staff and parents could communicate; those at home could view the material seen by the child at school and look at the activities that they had completed, and vice versa for the teacher, who could view what the child had done at home. To make the most of this potential to link MAPs together across locations it is also necessary to increase our understanding of these MAPs and how learners engage them in their learning.

For example, at the end of the study, Robert's mother reported that he very much enjoyed his numeracy work. She also reported that the amount that Robert talked about numeracy had increased from that before he had the Tablet PC and that his requests for help and her provision of assistance had also increased.

All parents whose children took part in the study were asked to complete a numeracy attitude questionnaire at the beginning of the study. Those that returned a completed questionnaire (n= 29) were sent a second, identical post-study questionnaire. Nineteen parents also completed both questionnaires and their post-study answers were compared

against their pre-study answers. Each question was asked against a 5 point Likert scale ranging from 'Not at all' to 'Very much / All the time'. One of the questions parents were asked was concerned with their child's requests for help. This is an important element of the learner' interactions with their more able partners, such as parents and other family members. These interactions are key to the Zone of Collaboration and to the Ecology of Resources model of context. This question was:

Q6. My child asks for help with their numeracy homework

The average parental answer to Q6 about help changed from pre to post study more than any other question and in a positive direction, indicating that children were asking for more help when doing their numeracy work with the HOMEWORK system than they had done previously. Parents also reported that children were choosing to work on their numeracy more without parents asking them, and in preference to other subjects. They reported that their children's interest had increased and that their children enjoyed their numeracy work.

One of the major aims of the HOMEWORK system development process was to build a model of the learner's interactions that could take into account their interactions across multiple locations and with multiple other people. In particular, the technology was developed to help link each learner's experiences at school with their experiences outside school, and to help provide conceptual coherence, so that the knowledge learnt at school was made relevant for home, too, and not seen as something that was only for formal school education. The system also extended the application of the scaffolding concept beyond the learner to explore the possibility of scaffolding parental interactions with their children, and to help family engagement and communication with the school.

The HOMEWORK system interface on the Tablet PC mapped out some of the resources that could be accessed through the Tablet PC by a learner and those helping that learner. Figure 5 illustrates this and shows that the resources available are categorized as those that arise from the child's interactions when at school, those that represent the history of that child's interactions inside and outside school, activities that are part of the Homework set by the teacher and fun activities. There is also a messaging facility for communications with the teacher. These represent the ZAA resources offered by the HOMEWORK system software through the Tablet PC. Interactions between the learner and other resources, such as the learner's family, and features of their environment can also be supported through the Tablet PC, as illustrated in the description of Robert's interactions.

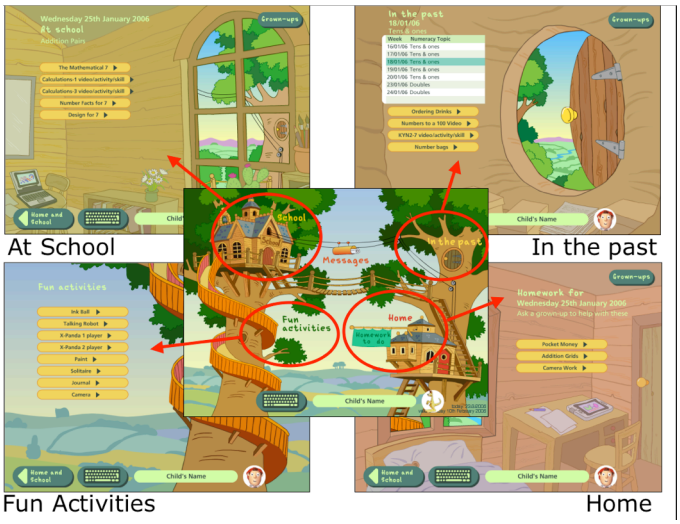


Figure 5 The HOMEWORK system interface when the tablet is out of range of the classroom network

As we stressed in earlier discussions about the role of the More Able Partner, the process of selecting resources to enable the construction of a ZPA is a negotiation between learners and More Able Partners. In a situation such as that described by the Homework case study, it is clear that in the out-of-school environment there may be various people who play the role of the More Able Partner at different points during the learner's interactions. Likewise, in the school environment, there will be the teacher, classroom assistant, peers, other teachers and parent helpers, each of whom may also fulfill the role of the More Able Partner at different points in time. The HOMEWORK system was designed to support both the child's learning and to support those in the position of the child's More Able Partner. It also had a part to play in the negotiation of the learner's ZPA with, and between, those playing the role of the child's More Able Partner, through, for example, the provision of information for parents and teachers about what each had done with the learner at home and at school. The ability to replay and review completed activities also offers each person acting as the learner's More Able Partner the ability to see what the child has done when either working alone or with another.

This emphasis upon the interactions between the different resources that a learner encounters is at the heart of the Ecology of Resources model that was used in the analysis of the data. Figure 6 illustrates the Ecology of Resources model for an extract of Robert's experience with the HOMEWORK system.

V. CONCLUSION

The Homework project has been used in this article as an example case study to demonstrate the empirical grounding of the Ecology of Resources model. The data from the project was also used to develop guidelines for the use of technology to support parental engagement. In

particular, the data and findings was combined with findings from another large study in the UK and reported in [25]. These includes, for example, highlighting the need for:

"- Carefully designed, parent focused support.
- Understanding what parents really need in order to help them get involved.
- Ensuring that continuity between in school and outside school is built, e.g. through carefully designed activities that aim to make work done at school relevant to the home context."

[25]

The findings presented in this paper draw upon fresh data examples and extend the discussions previously published in [26].

The definition of context that is discussed in this paper recognizes the interconnectedness of all the elements with which learners interact and the way in which these interactions shape our understanding of the world. Context should be considered as something that is defined with respect to an individual person:

> "it spans their life. A person's context is made up of the billions of interactions that they have with the resources of the world: other people, artefacts and their environment. These resources provide 'partial descriptions of the world' with which the learner can build connections through their interactions. These interactions help the learner to build an understanding of the world that is distributed across both resources and interactions: a distributed understanding that is crystallized with respect to a particular individual through a process of internalization."
>
> [5]

The Ecology of Resources model is offered as an abstraction that represents part of this reality for a learner, an abstraction that can be shared between social and technical researchers and practitioners to support analysis and to generate system design. It is concerned with learning and considers the resources with which an individual interacts as potential forms of assistance that can help that individual to learn. These forms of assistance are categorized as being to do with Knowledge and Skills, Tools and People and the Environment. These categories are not fixed, but rather offer a useful way of thinking about the resources with which a learner may interact and the potential assistance that these resources may offer. This emphasis upon the potential assistance that resources might offer highlights that it is the role that a particular resource element plays that is important, rather than its particularity. This emphasis upon context and the roles played by resources elements, including technologies, and upon the interactions between these resource elements means that the Ecology of Resources model could act as an integrative approach for Mobile Hybrid and On-line Learning. In this way the focus of the design process highlights the manner in which the resources and the relationships between them can be scaffolded and adjusted in order to meet the needs of the learner and to form their Zone of Available Assistance (ZAA).

The Ecology of Resources model is the basis for a design framework that offers a structured process through which educators and technologists can develop technologies and technology-rich learning activities that take a learner's wider context into account. This offers a basis upon which Distributed Scaffolding can be built. The process is participatory and iterative (for full detail see [5]).

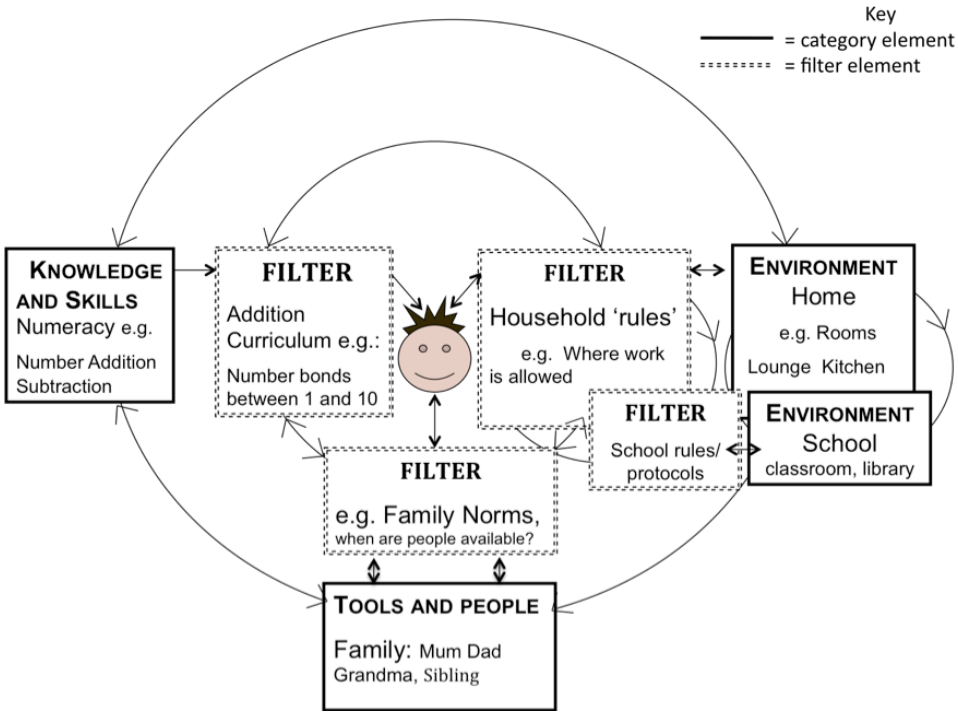


Figure 6 An example of an Ecology of Resources model of Robert's interactions with the HOMEWORK system.

VI.

VII. ACKNOWLEDGEMENT

The Ecology of Resources has been developed through a number of empirical studies to which many others have contributed. In particular, we thank Joshua Underwood, Wilma Clark and Fred Garnett for their insightful and constructive thoughts and comments. The Homework project used in the example in this paper is indebted to the teachers, parents and children for their help with this study that was funded by an EPSRC/ESRC/DTI PACCIT grant number RES-328-25-0027. We would also like to thank Joshua Underwood, Benedict du Boulay, Joe Holmberg, Lucinda Kerawalla, Jeanette O'Connor, Hilary Smith, Hilary Tunley and Catherine Luckin for their work on this project. Please note that children's names have been changed in order to protect their anonymity.

VIII. REFERENCES

[1] Luckin, 2010. The Ecology of Resources Model of Context: A Unifying Representation for the Development of Learning-oriented Technologies. Paper presented Paper presented at eKNOW 2010, The International Conference on Mobile, Hybrid, and On-line Learning. February 2010, Sint Maarten.

[2] Mercer, N. (1992). Culture, context and the construction of knowledge in the classroom. In: Light, P. & Butterworth, G. (eds.) Context and Cognition: Ways of Learning and Knowing, Mahwah, NJ, Lawrence Erlbaum, pp. 28-46.

[3] Wood, D., Underwood, J. & Avis, P. (1999) Integrated learning Systems in the Classroom. Computers and Education, 33 (2/3), 91-108.

[4] Nardi, B. (1996) Studying Context: A Comparison of Activity Theory, Situated Action Models and Distributed Cognition. In: B. A. Nardi (ed.) Context and Consciousness. Activity Theory and Human-computer Interaction. Cambridge, MA, MIT Press, pp. 69-102.

[5] Luckin, R. (2010) Re-designing Learning Contexts: technology rich, learner-centered ecologies. London, Routledge.

[6] Manovich, L. (2006) The Poetics of Augmented Space. Visual Communication, 5 (2), 219-240.

[7] Vygotsky, L. S. (1986) Thought and Language. Cambridge, MA, MIT Press.

[8] Vygotsky, L. S. (1978) *Mind in Society: The Development of Higher Psychological Processes*. Trans. Cole, M., John-Steiner, V., Scribner, S. &

Souberman, E. Cambridge, MA, Harvard University Press.
[9] Wood, D. J., Bruner, J. S. & Ross, G. (1976) The Role of Tutoring in Problem Solving. Journal of Child Psychology and Psychiatry, 17 (2), 89-100.
[10] Bruner, J. (1984) Vygotsky's Zone of Proximal Development: The Hidden Agenda. In: Rogoff, B. & Wertsch, J. V. (eds.), *Children's Learning in the 'Zone of Proximal Development'*. San Francisco, Jossey-Bass, pp. 93-97.
[11] Saxe, G. B., Gearhart, M. & Guberman, S. R. (1984). The social Organisation of early Number Development. In: Rogoff, B. & Wertsch, J. V. (eds.) *Children's Learning in the 'Zone of Proximal Development'*. San Francisco, Jossey-Bass, pp. 19-30.
[12] Valsiner, J. (1984) Construction of the Zone of Proximal Development in adult-child joint Action: The Socialisation of Meals. In: Rogoff, B. & Wertsch, J. V. (eds.) *Children's Learning in the 'Zone of Proximal Development'*. San Francisco, Jossey-Bass, pp. 65-76.
[13] Wood, D. J., Bruner, J. S. & Ross, G. (1976) The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry,* 17 (2), 89-100.
[14] Wertsch, J. V. (1984) The Zone of Proximal Development: Some conceptual Issues. In: Rogoff, B. & Wertsch, J. V. (eds.) Children's Learning in the 'Zone of Proximal Development'. San Francisco, Jossey-Bass, pp. 7-18.
[15] Wood, D. J. (1980) Teaching the young Child: Some Relationships between social Interaction, Language and Thought. In: Olson, D. (ed) *Social Foundations of language and cognition: Essays in Honor of J.S. Bruner.* New York, Norton.
[16] Wood, D., Shadbolt, N., Reichgelt, H., Wood, H. & Paskiewitz, T. (1992) EXPLAIN: Experiments in planning and instruction. *Society for the Study of Artificial Intelligence and Simulation of Behaviour Quarterly Newsletter,* 81, 13-16.
[17] Koedinger, K. R., Anderson, J. R., Hadley, W. H. & Mark, M. A. (1997) Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43
[18] Li, D. & Lim, C. (2008) Scaffolding online historical inquiry tasks: A case study of two secondary school classrooms. *Computers and Education*, 50, (4), 1394-1410
[19] Nussbaum M., Alvarez, C., McFarlane, A., Gomez, F., Claro, S. & Radovic, D. (2009) Technology as small group face-to-face Collaborative Scaffolding, *Computers & Education,* 52, 147-53
[20] Azevedo, R., Moos, D.C., Winters, F.W., Greene, J.A., Cromley, J.G., Olson, E.D. & Godbole-Chaudhuri, P. (2005) Why is externally-regulated learning more effective than self-regulated learning with hypermedia? In: Looi, C-K. McCalla, G. Bredeweg, B. & Breuker, J. (eds.) *Artificial Intelligence in Education: Supporting Learning through intelligent and socially informed Technology.* Amsterdam, Netherlands, IOS Press, pp. 41-48.
[21] Wood, H.A. & Wood, D. (1999) Help seeking, learning and contingent tutoring. *Computers and Education*, 33, 2–3, 153-69.
[22] Puntambekar, S. & Kolodner, J. L. (2005). Distributed scaffolding: Helping students learn science by design. *Journal of Research in Science Teaching*, 42, (2), 185-217
[23] Tabak, I. (2004) Synergy: A complement to emerging patterns. *Journal of the Learning Sciences,* 13, (3), 305-35.
[24] Luckin, R., du Boulay, B., Underwood, J., Holmberg, J., Kerawalla, L., O'Connor, J., Smith, H. and Tunley, H. (2006) Designing Educational Systems Fit for Use: A Case Study in the Application of Human Centred Design for AIED. International Journal of Artificial Intelligence in Education, 16, 353-80.
[25] Lewin, C. & Luckin, R. (2010) Technology to support parental engagement in elementary education: Lessons learned from the UK . Computers and Education Vol 54, Issue 3, Pages 749-758
[26] Luckin, R. (2008) The learner-centric Ecology of Resources: A Framework for using Technology to scaffold Learning. Computers & Education, 50, 449-62.

Africa's Telenursing Today (and Tomorrow?)

Siinclair Wynchank, Jill Fortuin
Telemedicine Platform
Medical Research Council
Cape Town, South Africa
SWynchank@mrc.ac.za; Jill.Fortuin@mrc.ac.za

Abstract— **Although telenursing is well established in developed nations, developing nations have much less such activity. In Africa, and particularly South Africa, some effective pilot telenursing schemes have been introduced and are presented here. A further success has been in a wide variety of teleeducation programmes which target nurses. Usually telenursing equipment and procedures that have been successful in the well-resourced world are inappropriate for transfer, without modification, to developing countries. However African telenursing applications have used relevant prior knowledge in the field, wherever possible. These applications, with a likely future course and means of accomplishing it, will be outlined using what little quantitative data is available.**

Keywords - Telenursing; Africa; Telemedicine; Teleeducation; Distance learning.

I. INTRODUCTION

Because of a general shortage of physicians in Africa, its nurses often have greater responsibilities than in many other countries and much of the continent's public health care is provided by nurse practitioners in village clinics. Such nurse-directed clinics, also found in some small towns, form the backbone of African primary health care (PHC) services. These facilities are commonly far from the nearest medical doctor, or hospital, and they are usually located in regions where local infrastructure and transport services are poor. So, African nurses often have much authority. Telenursing can greatly aid their provision of PHC, and their management of both chronic conditions and certain acute situations. But there are significant differences between telenursing in developed countries and Africa, because of funding available, experience with, and availability of, information and communication technology (ICT), other infrastructure and physical conditions, the nature of their patients' illnesses, stage of presentation of the diseases encountered, etc. Many of these differences are also applicable to telenursing in other poorly-resourced regions.

Telenursing has many definitions, but we consider it as the use of ICT and electronic transfer of information relevant to nursing. As we shall show, telenursing can support virtually all aspects of nursing activity, often in ways not otherwise possible. This is especially true in Africa, where personnel, equipment and expertise often lack, as is the case sometimes in the developed world's most remote rural situations. In this report a nurse is deemed someone, often a female, who has been trained to care for those experiencing illhealth; the sick, infirm and disabled. Equally importantly, a nurse promotes healthy growth and development in children and the establishment and maintenance of health in all age groups. Telenursing is closely associated with telemedicine [1], which has been available since before the invention of the telephone, whose availability greatly increased telenursing and telemedicine activity in the first half of the last century. However telenursing has greatly increased in the last few decades. This results from the recent vast explosion of computer power and the wide variety of readily available electronic communication modalities, whose costs are steadily decreasing.

From a search of the U.S. National Library of Medicine's database it was determined that there are over 150 times more publications on telenursing from developed countries than on African telenursing, which today is in its infancy [2]. Hence its practices, norms and applications are still in the process of being established. Because application of ICT is necessary for telenursing, it was first set up where there was appropriate funding, technical expertise and infrastructure. Therefore in developed countries, telenursing grew steadily and, after its utility was confirmed, it began to be applied to the needs of the poorest nations in Africa and other developing regions.

Today's level of African telenursing justifies a critical study of its history, current applications and impact. This study is based mainly on South African telenursing. Using South African experience is particularly apposite now, for several important reasons. In contemporary rural, and much of peri-urban, South Africa there are social and economic conditions that are characteristic of the poorest nations of Africa. (In contrast, there are also infrastructure, nursing and medical services in South Africa of a high standard, principally in the private health care domain of large urban centres.) Before democracy was established in 1994, South Africa was almost totally isolated from other African nations. Since that date warm relations have developed between South Africa and other nations of the continent, resulting in steadily increasing numbers of

cooperative projects in all medical fields and throughout Africa. Such health care endeavours have increasingly allowed South Africa to use its expertise and experience, gained in its own poorly developed remote rural regions, for the benefit of others. Also experience gathered in well-resourced nations is relevant to Africa's telenursing, for lessons learned elsewhere can reduce wasteful duplication. At the same time it is essential to appreciate that often telenursing practice, training, equipment and application in a wealthy country may be inappropriate to be transferred without change to a public health service in Africa. Here the physical and other circumstances and needs often differ from those of the developed world. So for effective introduction of telenursing, modification is frequently essential.

There is a world-wide lack of all types of health care workers (HCWs), especially so in the most economically-deprived nations, many of which are found in sub-Saharan Africa. [3] Nurses are usually the principal HCWs in poorly-resourced nations, especially in their rural regions. Yet the attitudes of patients to nursing practitioners in developing countries have not been extensively investigated, in contrast to the USA. There such nurses, with more independence and clinical responsibility, have been practising for over 40 years and their professional activity is known to the vast majority of the USA population. They are in great demand and about 60% of that nation's population have used their services. A large majority (82%) of their patients were satisfied, or better, with their ministrations. [4] Similar or better results are probable in Africa, where fewer health care alternatives are available. In South Africa most public health service rural clinics and dispensaries are located in villages, with rare visits (or often no visits at all) from medical doctors. Thus, the potential for telenursing to support and improve these nurse practitioners' service to their local public is very great. Another aspect of telenursing, particularly valuable for Africa, is provision of continuing professional development using distance learning. This important topic is already the subject of current successful projects and will be considered in greater detail below.

Worldwide there are few formal qualifications concerning the practice of telenursing and in most African countries very little incorporation of telenursing and associated subjects in the state-approved nursing curriculum. Although South Africa has many, varied, ongoing telenursing activities, the South African Nursing Council, a statutory body, which, "controls and exercises authority in respect of all matters affecting the education and training of registered nurses, midwives, enrolled nurses and enrolled nursing auxiliaries" and determines nursing curricula, has no material on telenursing in the official curriculum of 2010, although ICTs do feature. So in South Africa and most other countries telenursing is usually informally taught with few recognised requirements for its practitioners. However in Queensland, Australia, the State Government has introduced the formal qualification of "Telenurse". This can only be obtained by a state-registered nurse, who must then undergo further study and training. [5] Most persons involved in telenursing (both telenurses and callers) are female, so gender issues can play an important role in telenursing. A telenurse must be able to interact with a caller in a wide variety of ways. This is because on occasion the telenurse may be required to advise, assess, refer, support, teach and also offer selfcare advice and triage, if necessary. [6] Therefore telenursing requires a marked level of competence. African state-registered nurses have this in abundance, for often they have more responsibility than their equivalents in developed countries and their professional instruction is of a high standard. This is confirmed by the ongoing brain drain of African state-registered nurses to European and other developed nations and the frequent lack of examinations that they must pass before they are allowed to practise after emigration.

A shortage of HCWs exists in all developing countries and detailed World Health Organisation (WHO) studies of HCW needs in Ethiopia, Ghana, Kenya, Malawi and Tanzania suggest approaches to improve this situation. These include increased use of modular education and ICTs. [7] Telenursing can also play a major role in such programmes in all African developing countries. There, over 60% of health problems result from largely preventable infections (e.g., malaria, tuberculosis (TB) pneumonias and viral infections [8]). Adaptations necessary for African telenursing to be practical and effective and the roles of educational and clinical telenursing activity will be outlined and discussed below.

II. AFRICAN TELENURSING'S STATE OF THE ART

Recent publications have indicated that in the last year there have been many significant advances in African telenursing. However since infrastructure is often lacking, simple modalities of communication are most often employed. In developing countries about half of reported telemedicine uses Email, with many fewer real time interactions. [9] Mobile phones are readily available .throughout Africa and their utility to support management of the chronically ill of the poorest Hondurans has been reported. [10] Moves are afoot to reproduce this in Africa. A successful pilot scheme in Malawi combines mobile phones and text messages to contact patients up to 100 miles from the base health care facility. [11] After 6 months, 2000 hours of travel time were saved and the number of patients in an anti-TB programme doubled. Cervical carcinoma is very common in Africa and a novel telenursing network attempts "to bridge the gap between screening and diagnosis" in Zambia. [12] Use of digital cervicograms allows participating nurses to discuss their significance with patients and permits further consultation as required. But in spite of these successes, it is very clear that "The human touch is essential. A named health care provider with access to telehealth" facilities must, wherever possible, be available. [13] Incidence of

cardiac problems is steadily increasing in Africa and it has been clearly demonstrated that a telenursing programme directed towards aiding chronic heart failure results in "improved patient care and medication concordance and reduced use of health care services". [14] Similar advantages have been reported for preventive care where telenursing intervention improved cardiovascular disease risk awareness. [15] This work presents the current situation of African telenursing and some indications of how it may evolve.

III. TELEEDUCATION IN AFRICA FOR NURSES

Teleeducation applications of telemedicine in Africa have demonstrated even more success, than clinical health care. [16] Although telenursing links are usually set up to transmit queries, requiring a consultant's response, and information concerning patient healthcare, their bidirectional nature also enables educational material to be sent to nurses at telenursing sending locations. African nurses have enthusiastically embraced such teleeducation. It can provide both theoretical and practical knowledge. Also it can allow the sharing of experience to improve the quality of professional services offered to patients. Conventional face-to-face, teaching methods, whether residential or not, can achieve the same ends, but these may involve travel (often over large distances for rural African nurses), significant time spent away from the usual workplace and accommodation expenses. Teleeducation can overcome many of these limitations and costs. Teleeducation for nurses in Africa can not only provide possible future theoretical educational possibilities. It is important to recognise that there have already been concrete achievements at a variety of levels in African distance learning for nurses.

Web-based distance learning is now possible for all countries with Internet connections and according to UNESCO (the United Nations Educational, Scientific and Cultural Organisation) there are at least 53 such African nations. [17] Its success has been clearly demonstrated in several existing African nursing projects, which are web-based, sometimes totally so. For example nurses usually play a crucial role in palliative care. The only recognised qualifications obtainable in Africa in this field of study are a postgraduate diploma and degree (M Phil), both offered by the University of Cape Town, South Africa. These courses depend almost entirely on web-based distance learning. Since the course's beginning, 11 years ago, it has trained over 200 students from 12 English speaking African nations. A very recent educational evaluation of this programme considered the 125 registrants of the course, who were enrolled between 2000 and 2007. [18] It concluded that it is, "applicable to current community needs and appropriate for the participants". Further, the assessment expects that, "palliative and hospice care no doubt will evolve into a mainstream service within South Africa's health care system". Such are the impressive consequences of one specific African teleeducational course. This postgraduate programme is only open to nurses and medical doctors. Its curricula are adjusted for conditions in the student's homeland and emphasise family dynamics, both for extended and nuclear families, since they must have much relevance for the current scourge of HIV/AIDS that is now sweeping across sub-Saharan Africa. Although there is a clear need for these skills in Africa, many of the continent's countries have no expertise in palliative care. [19] In 2011, the same university will offer another Master's degree course in Public Mental Health, based on teleeducation. This too will be directed towards postgraduate nurses and in its first year will emphasise planning and implementation of policy. [20] The University of Cape Town has also made available 14 modular, computer-based teleeducation courses on community paediatrics. These are intended for nurses in rural community health centres and clinics, who often have no direct access to paediatricians. [21] Teleeducation in other nursing fields is also available and/or under development. In South Africa there is compulsory community service of one year for newly graduated nurses, who are usually assigned to remote rural areas where there is the greatest lack of HCWs. In spite of working with experienced colleagues, they frequently feel very isolated. With telenursing facilities at their workplace this sense of isolation is greatly reduced, leading to a happier and more productive community service, in addition to valuable acquisition of knowledge.

Moroccan nurses have available a paediatric oncology-haematology web-based programme that also serves nurses in other developing countries. [22] Overall, paediatric cancer cure rates are about 75% in developed countries and a third of this in developing nations, where it can be the leading cause of death in children aged 5-15 years. Hence such a programme can save many lives, for about 80 - 85% of all childhood cancers occur in developing countries, where survival can be under 10%. [23] Another application of a web-based site has been for a nursing study in Ethiopia, and other developing countries, investigating a mother's mental health and her child's nutrition status. This work "confirms that promotion of maternal mental health may be important for the improvement of child nutrition." [24] Obstetric complications are a frequent cause of death in developing countries and worldwide this causes >500,000 deaths yearly. [25] The assessment and improving of skilled birth attendants in Benin, Rwanda and other developing countries has recently been underway, with WHO support. A website played an important role in training local assessors and this successful pilot scheme has been extended to Niger and Kenya. [26] A joint Eritrean-USA teleeducation programme for nurses provides instruction in challenging aspects relevant to their practice in Eritrea. One of its aims is to train tutors for local nursing schools. [27] Midwifery was selected as the first discipline for this approach. Appropriate technology was introduced and the

overall emphasis was on clinical application. Student input and flexibility were crucial parts of the curriculum design.

A particularly distressing statistic is that of recent trends in infant mortality rates. In many developing countries they have either remained static or have increased (as in South Africa) during the last few years. Such deaths are mainly preventable and most often are caused by infectious diseases (pneumonias, tetanus, etc) in the first week of life. Such mortality can be reduced by appropriate and sufficient nutrition, particularly breast feeding. This assists in giving the infants antibodies to fight infectious diseases, as well as nourishment. UNICEF, the WHO and other bodies recognise that nurses can play an important role in improving this situation. So an "essential newborn care programme" to decrease high infant mortality rates was drawn up and has been evaluated. Use of a self administered, computer based course was found to be equally successful in increasing relevant knowledge as the conventional face to face course in South African and Zambian studies. [28] In developing countries, resources are lacking and the lower cost of a teleeducation course in this critical field can allow much greater dissemination of relevant knowledge, for a given level of funding.

In order to benefit from telenursing advances in developed nations, without duplication of the efforts required to make these advances, it is essential to examine how established telenursing programmes, and those under development, can be adapted for use in Africa. Some examples follow. Disaster aid in regions affected by abrupt deterioration/destruction of health care facilities or sudden and overwhelming need for such facilities is often provided by nurses. A British postgraduate qualification in nursing for post disaster relief uses teleeducation and emphasises aspects of transcultural nursing. All these characteristics make it especially useful in the African context. [29] In the Netherlands nurses drive a pilot programme to enable appropriate and vulnerable patients to monitor, and if necessary change, their risk profile for vascular disease, by teaching them to self-manage more effectively. [30] This interactive principle is suitable for some African applications, for it is noted that vascular disease prevalence is steadily increasing in Africa, paralleling the continent's rapid urbanisation of the last few decades and the resulting life style changes.

Developing countries often lack proficiency in paediatric emergency care and the necessary facilities are scarce. A training scheme to address this lack in Vietnam was established initially in Australia for Vietnamese nurses with different professional backgrounds, and others. It also undertook instructor training and provided organisational experience. [31] This training programme proved sustainable for transfer to the developing regions in need of this expertise. African nations can clearly benefit from this approach, to obtain improved expertise in this and other fields and moves are under way to establish such a course in Africa. The programme and necessary updates use telenursing techniques.

In Malawi, as elsewhere in Africa, malaria is a principal cause of morbidity and mortality, especially for children, in whom about 90% of the life threatening form occurs. The greatest mortality is found in those less than 5 years of age. Relevant clinical management in Malawi was studied [32] and the principal result was that the care offered to children at the first referral level, principally given by nurses, required revision. Telenursing can provide the necessary information to those in PHC clinics to implement this important finding, in the many countries of Africa where malaria is a serious health problem.

In South Africa a video programme, transmitted from a central location and operating for 4 years, is directed by nursing staff for the benefit of those in waiting rooms of public health service facilities. This programme aims to improve the patients' understanding of HIV/AIDS and other conditions. Results are very encouraging and it is to be extended to neighbouring nations. [33]

IV. TELENURSING'S ADAPTATIONS FOR AFRICA

African public health budgets are usually meagre, so extensions or modifications to existing telenursing equipment must be carefully evaluated to prevent waste of resources. The simplest procedure for telenursing, to purchase equipment and then connect it to an existing communications network, is often inappropriate for Africa. [34] So before embarking on an African telenursing project the complete project and its integration with available facilities should be very carefully reviewed. Obtaining the supportive involvement of central government, local medical bodies and leaders is crucial, especially for questions of regulation and incorporation with pre-existing health services. [35] The technology and procedures finally selected will depend on the users' computer literacy abilities, available infrastructure, etc. All these points may be self-evident, but they are too often passed over for many reasons, ranging from a lack of technical knowledge, to equipment sellers' lack of scruples. Because of incoordination between the nine South African provinces' ICT standards, there is no single ICT system for HIV/AIDS management that can be used in the whole country. [36] Hence, adequate preparation before setting up a telenursing network is essential. If possible, nurse users should be invited to use the equipment, before a final choice is made. Frequently their suggested modifications in procedures, equipment, etc., may be of great benefit, for this can improve network and other function. [37] As an example of such benefits, a ruggedised, simple, telenursing workstation, designed in South Africa, was modified by adding a remote control and simplified menus. So then it became much more acceptable for its nurse users, whose previous ICT experience was minimal or zero.

Cellular/mobile phone systems have been introduced into almost all African countries and adopted so

enthusiastically at all levels of society that 40% of Africans have a mobile phone. [38] Since many poor regions of African nations have satisfactory phone signals, this technology can be incorporated into African telenursing, as has been successfully done in India. [39] Such application has been beneficial for management of serious chronic conditions [40] and to ensure appropriate action in medical emergencies, such as those occurring in childbirth. The latter suggestion was made for Burkina Faso. [41] What is now new in developed nations' telenursing will later often become available, if required, in Africa. But local African needs and conditions must be considered to ensure the most effective application. For example software used to aid decision making by telenurses has both pros and cons. It simplifies telenursing and often complements telenurses' knowledge. But also it can be incomplete and in conflict with their independently formed decisions. Overall telenurses preferred having it available, although it cannot always replace their own nursing expertise. [42] Such software, being considered for use in Africa, should be modified, for example to include and emphasise conditions common in the continent, but much rarer elsewhere.

Kenya is the only sub-Saharan African developing nation with a nursing database. [43] It provides reliable nursing information about workforce capacity, demographics and migration patterns, so allowing assistance in determining optimum distribution and most effective use of health personnel.

V. AFRICAN CLINICAL TELENURSING

African telenursing has been able to benefit from experience in developed nations where telenursing has been applied and steadily developed. Resulting knowledge and experience gained there have been adapted and expanded as necessary to provide telenursing most suitable for Africa. Transfer of such activity and knowledge to Africa has accelerated as costs of equipment have decreased and ICT availability and expertise have both improved in recent years.

It is well known that HIV/AIDS is currently savaging Africa. About 60% of the world's HIV infections are in sub-Saharan Africa [44] and the world's largest current anti retroviral therapy programme is in South Africa. [45] Applications of telenursing can alleviate many aspects of this scourge. Feeding of infants by mothers infected by HIV/AIDS has been studied by the WHO and guide-lines drawn up to aid prevention of virus transmission from the mother to infant. Three clear conditions for this feeding have been established, to greatly reduce the occurrence of infant infection. However, over two thirds of South African HIV/AIDS infected mothers do not fulfil these criteria, with increased risk of viral transmission to their infants. Appropriate counselling of these mothers at their baby clinics, or elsewhere, is urgently required, not only in South Africa but in many other similarly affected African nations. Telenursing links can facilitate provision of guidelines for nurses to provide such counselling and also to make available counsellor training. [46]

Stigma of those infected by HIV remains a serious problem in many African societies and this has been unambiguously shown to violate human rights in South Africa, Swaziland, Tanzania, Lesotho and Malawi. [47] Until this stigma is overcome and the associated damage redressed many believe that the pandemic will continue, because of the consequent difficulty in ensuring that individuals' HIV status known. A recent detailed survey in Tanzania, Zimbabwe and South Africa, indicates that education and widespread HIV testing are essential to produce any significant reduction of HIV related stigma. [48] A teleeducation programme, outlined above, attempts to provide such education. [33] A report has shown that Zambian nurses, either infected with HIV/AIDS or caring for HIV positive patients, can be greatly aided by participating in local support groups, which focus on appropriate training and monitoring. [49] Telenursing techniques allow such assistance to be rendered more effectively at reduced cost, than by using traditional methods.

A South African project illustrates how an initial emphasis on clinical consultation via telenursing can lead to improved nursing ability and service to the public in several different ways, which include teleeducation. At a nurse-directed clinic, in a poor region where there is much alcoholism and violence, disturbed patients often present. This provides a dilemma, especially at the weekend, when specialist opinion is unavailable. The nurses have had little training in psychology and the only management available at the weekend is for the patient to be transferred to police cells, to await the arrival, on the following Monday, of the district surgeon. S/he likewise may have had little psychological training. A department of psychology (of the University of the Western Cape, in South Africa), located in the provincial capital, is responsible for this clinic's mental health practice, but due to its location 450 km away, there is infrequent direct contact. An audio-visual telepsychology link greatly improved this situation and other benefits quickly followed. Psychology students in the university could experience pathology, rarely encountered in their urban setting, through the telenursing link. Practical training for the clinic nurses in managing disturbed, and other, patients became possible. Also two other distance learning programmes, provided by staff members of the department of psychology and using the telenursing link, resulted in much additional benefit. Both programmes trained volunteer lay counsellors. One was directed towards local professionals, such as clergymen, librarians, school teachers, social workers, etc. The other novel programme taught high school pupils, as peer-counsellors and this proved particularly effective. Such programmes would have been totally impractical without a teleeducation facility. It was much regretted that this programme ended abruptly because of equipment theft.

South African telenursing programmes are being extended throughout the continent under the aegis of regional bodies, such as the Southern African Development Community (SADC) with 15 southern African members and the New Partnership for African Development (NEPAD), with 18 nation members in all parts of Africa. Both SADC and NEPAD emphasise collaborative health programmes for African nations and actively encourage telemedicine.

DISCUSSION

It is clear that in Africa there are many telenursing activities underway; some are pilot schemes and others are well established. It is now appropriate to enlarge their scope and to involve more African countries. South Africa is ready to take a leading role in this extension, having had a head start, as shown by its current telenursing activities, some of which have been outlined above. African telenursing is changing rapidly, as is indicated by the recent publication dates (2007 and later) of about 80% of the references to this paper. Benefits of telenursing are well known and varied, including savings in travel costs and increases in nursing expertise. After several patients with similar conditions have been referred by a telenurse, subsequently s/he is often able to manage comparable patients without referral. This was clearly demonstrated by a local pilot dermatological telenursing project. The referring telenurses very often encountered HIV infected patients, for >95% of HIV infected patients have dermatological lesions and the severity of such dermatological pathology closely reflects progress of the HIV infection. After the telenursing link provided management regimes, for a few weeks the participating nurses' steadily increasing experience enabled them to diagnose and manage increasing numbers of HIV lesions without referral.

Very little analysis of economic benefits of African telenursing has been performed. However in developed countries diminished travel costs provide clear savings and this benefit is becoming apparent in Africa too. The telenurses' increased experience from telereferrals has already been outlined and its consequence is an improved local health care service. The principal advantages of teleeducation are plain, for the resultant nursing education of all sorts, especially continuing professional learning, is less costly when travel is eliminated. African telenursing is evolving to satisfy African needs. It recognises that technical and infrastructure conditions and pathology in the continent may differ from those of the nations where telenursing equipment and applications were first derived, so for an optimal role, modifications are often necessary, as summarised above.

Overall the essential benefit of telenursing is an improved service for those whom the nurses serve, especially patients receiving PHC. This has been recently confirmed in an extensive pilot programme in Brazil, which is intended to a model for a National Telehealth Program. [50] Its findings substantiated South Africa's related experience, which is that both direct PHC health care and the professional knowledge of the nurses who administer it, benefit significantly. A report on this pilot found that "Telehealth has strengthened the role of primary health care… It has reinforced primary care units … and has provided primary care staff with a powerful arsenal of up-to-date information and tools". [50] Telenursing is usually less expensive than traditional nursing, but there are situations when it may even be more costly, for in some rural areas of Africa it replaces a situation where no local service at all was previously available. Today, all Australasians have free access to nursing advice via the phone. [51] One day such telenursing service will be available in Africa. The path to this goal is long, but the first steps of this journey have been made, by preparing projects using the rapidly increasing numbers of mobile phones in Africa.

CONCLUSION

Telenursing activity in Africa has steadily increased, as the work of telenursing pioneers, in well-resourced nations has been studied and adapted according to African needs, conditions and infrastructure. It is essential to modify techniques and equipment in terms of the continent's technical, infrastructural and computer literacy levels and its different range of pathologies. This can ensure effective application of telenursing in Africa. Also teleeducation has greatly benefited African nurses and, with the wide range of existing collaborations, it seems that Africa's particular form of telenursing is now set to expand steadily, so that it will continue to contribute significantly to marked improvements in Africa's health care.

ACKNOWLEDGMENTS

The authors wish to thank Professors A. Flisher, E. Gwyther and M. Kibel for informative discussions.

REFERENCES

[1] S. Wynchank and J. Fortuin, "African Telenursing: what is it and what's special about it?", eTELEMED 2010, 2nd Int Conf on eHealth, Telemedicine, and Social Medicine. Feb 10-15, 2010. St. Maarten, Netherlands Antilles, pp 17-22.

[2] Priv. comm., 2010, A. Wynchank.

[3] L. Ogilvie, J.E. Mill, B. Astle, A. Fanning, and M. Opare, "The exodus of health professionals from sub-Saharan Africa: balancing human rights and societal needs in the twenty-first century", Nurs Inq, 2007 Jun; vol 14(2): pp 114-24.

[4] D.J. Brown, "Consumer perspectives on nurse practitioners and independent practice", J Am Acad Nurse Pract. 2007 Oct; vol 19(10): pp 523-9.

[5] Anonymous, "Professional qualification of 'telenurse' (Queensland, Australia)", 2009 from website: http://www.health.qld.gov.au/workforus/careers/Telenurse.pdf, Dec 2010.

[6] E. Kaminsky, U. Rosenqvist, and I. Holmström, "Telenurses' understanding of work: detective or educator?", J Adv Nurs. 2009 Feb;65(2): pp 382-90.

[7] Anonymous, "WHO. World Health Report 2006: working together for health", World Health Organisation, Geneva, Switzerland, 2006

[8] N. Crisp, B. Gawanas, and I. Sharp, "Training the health workforce: scaling up, saving lives", Lancet. 2008 Feb 23; vol 371(9613): pp 689-91.

[9] R. Wootton R and L. Bonnardot," In what circumstances is telemedicine appropriate in the developing world?", JRSM Short Rep, 2010 Oct 1;1(5):37.

[10] G.E. Quinn, C. Gilbert, B.A. Darlow. and A Zin, "Retinopathy of prematurity: an epidemic in the making", Chin Med J (Engl). 2010 Oct;123(20):2929-37.

[11] N. Mahmud, J. Rodriguez, and J. Nesbit, "A text message-based intervention to bridge the healthcare communication gap in the rural developing world", Technol Health Care 2010 Jan;18(2):137-44

[12] G.P. Parham, et al., "eC3--a modern telecommunications matrix for cervical cancer prevention in Zambia", J Low Genit Tract Dis, 2010 Jul;14(3):167-73.

.[13] I. Holmström, "Diabetes telehealth and computerized decision support systems: a sound system with a human touch is needed", J Diabetes Sci Technol, 2010 Jul 1;4(4):1012-5.

[14] R. Berkley, S.A. Bauer, and C. Rowland, "How telehealth can increase the effectiveness of chronic heart failure management", Nurs Times. 2010 Jul 6-12;106(26):14-5.

[15] L. Jensen, et al., "Impact of a nurse telephone intervention among high-cardiovascular-risk, health fair participants", J Cardiovasc Nurs, 2009 Nov-Dec;24(6):447-53.

[16] M. Mars, "Health capacity development through telemedicine in Africa", Yearb Med Inform. 2010:87-93.

[17] J. Anderson, "ICT transforming education; a regional guide", 2010, UNESCO, Paris, France. ISBN 978-92-9223-325-9.

[18] C.D.L. Ens, et al., "Postgraduate palliative care education: Evaluation of a South African programme", S Afr Med J, 2011, Jan, 101(1) 42-44.

[19] Priv. comm., 2010, E. Gwyther.

[20] Priv. comm., 2010, A. Flisher.

[21] Priv. comm., 2010, M. Kibel.

[22] J.A. Wilimas and R.C. Ribiero, "Pediatric hematology-oncology outreach for developing countries", Hematol Oncol Clin North Am. 2001 Aug;15(4): pp 775-87.

[23] S.W. Day, P.M. Dycus, E.A. Chismark, and L. McKeon, "Quality assessment of pediatric oncology nursing care in a Central American country: findings, recommendations, and preliminary outcomes", Pediatr Nurs. 2008 Sep-Oct; vol 34(5): pp 367-73.

[24] T. Harpham, S. Huttly, M.J. De Silva, and T. Abramsky, "Maternal mental health and child nutritional status in four developing countries", J Epidemiol Community Health. 2005 Dec; vol 59(12): pp 1060-4.

[25] M. Islam, "The safe motherhood initiative and beyond", Bull World Health Organ, 2007, vol 85, p 735

[26] S.A. Harvey, et al., "Are skilled birth attendants really skilled? A measurement method, some disturbing results and a potential way forward", Bull World Health Organ, 2007, vol 85, p 783

[27] P. Johnson, G. Ghebreyohanes, V. Cunningham, D. Kutenplon, and O. Bouey, "Distance education to prepare nursing faculty in Eritrea: diffusion of an innovative model of midwifery education", J Midwifery Womens Health. 2007 Sep-Oct; vol 52(5): pp e37-41.

[28] E.M. McClure, et al., "Evaluation of the educational impact of the WHO Essential Newborn Care course in Zambia", Acta Paediatr. 2007 Aug; vol 96(8): pp 1135-8.

[29] K. Davies, P. Deeny, and M. Raikkonen, "A transcultural ethos underpinning curriculum development: a master's programme in disaster relief nursing", J Transcult Nurs. 2003 Oct;14(4): pp 349-57.

[30] B.M. Goessens, et al., "A pilot-study to identify the feasibility of an internet-based coaching programme for changing the vascular risk profile of high-risk patients", Patient Educ Couns. 2008 Oct;73(1): pp 67-72.

[31] S. Young, et al., "Teaching paediatric resuscitation skills in a developing country: introduction of the Advanced Paediatric Life Support course into Vietnam", Emerg Med Australas. 2008 Jun;20(3): pp 271-5.

[32] P.P. Diep, L. Lien, and J. Hofman, "A criteria-based clinical audit on the case-management of children presenting with malaria at Mangochi District Hospital, Malawi" World Hosp Health Serv. 2007; vol 43(2): pp 21-9.

[33] A. Deverell, et al., "An Evaluation of Mindset Health: Using ICT to facilitate an innovative training methodology for health care providers in South Africa" 11th Conf Int Soc Telemed eHealth, Cape Town, South Africa, Nov 2006.

[34] P. Yellowlees, "How not to develop telemedicine systems", Telemed Today 1997 May-Jun; vol 5(3): pp 6-7, 17.

[35] E.K. Yun and H.A. Park, "Factors affecting the implementation of telenursing in Korea", Stud Health Technol Inform. 2006;122: pp 657-9.

[36] T. Sørensen, U. Rivet, and J. Fortuin, "A review of ICT systems for HIV/AIDS and anti-retroviral treatment management in South Africa", J Telemed Telecare. 2008; vol 14(1) pp 37-41.

[37] D. Hibbert, et al., "Lessons from the implementation of a home telecare service", J Telemed Telecare. 2003;vol 9 Suppl 1: pp S55-6.

[38] Anon., "The power of mobile money", The Economist, 2009 Sep 26; vol 392(8650): p 13.

[39] G.R. Kanthraj and C.R. Srinivas, "Store and forward teledermatology", Indian J Dermatol Venereol Leprol. 2007 Jan-Feb; vol 73(1): pp 5-12.

[40] H. Blake, "Mobile phone technology in chronic disease management", Nurs Stand. 2008 Nov 26-Dec 2; vol 23(12): pp 43-6.

[41] P. Byass and L. D'Ambruoso, "Cellular telephone networks in developing countries", Lancet 2008 Feb 23; vol 371(9613): p 650.

[42] A. Ernesäter, I. Holmström, and M. Engström, "Telenurses' experiences of working with computerized decision support: supporting, inhibiting and quality improving", J Adv Nurs. 2009 May;65(5): pp 1074-83

[43] P.L. Riley, et al., "Developing a nursing database system in Kenya", Health Serv Res. 2007 Jun;42(3 Pt 2): pp 1389-405.

[44] Anon., "UNAIDS Report on the global AIDS Epidemic 2010", UNAIDS and WHO; Geneva, Switzerland: UNAIDS/WHO, 2010. ISBN 978-92-9173-871-7

[45] M. Cornell, et al., "Monitoring the South African National Antiviral Treatment Programme, 2003-2007: The IeDEA Southern Africa collaboration", 2009, S Afr Med J, vol 99, No 9, pp 653-660.

[46] T. Doherty, et al., "Effectiveness of the WHO/UNICEF guidelines on infant feeding for HIV-positive women: results from a prospective cohort study in South Africa", AIDS. 2007 Aug 20; vol 21(13): pp 1791-7.

[47] T.W. Kohi, et al., "HIV and AIDS stigma violates human rights in five African countries", Nurs Ethics. 2006 Jul;13(4): pp 404-15.

[48] B.L. Genberg, et al., "A comparison of HIV/AIDS-related stigma in four countries: negative attitudes and perceived acts of discrimination towards people living with HIV/AIDS", Soc Sci Med, 2009 Jun; vol 68(12): pp 2279-87.

[49] M.P. Vitols, E. du Plessis, and O. Ng'andu, "Mitigating the plight of HIV-infected and -affected nurses in Zambia", Int Nurs Rev. 2007 Dec; vol 54(4):pp 375-82.

[50] A. de Fàtima dos Santos, et al., "Telehealth in Primary Healthcare: An Analysis of Belo Horizonte's Experience", Telemed J E Health. 2011 Jan 7 vol 17(1): pp 25-9.

[51] I. St George, M. Cullen, L. Gardiner, and G. Karabatsos, "Universal telenursing triage in Australia and New Zealand - a new primary health service", Aust Fam Physician. 2008 Jun; vol 37(6): pp 476-9.

What Motivates Faculty to Adopt Distance Learning?

Data Collected from a Faculty Development Workshop Called "Build a Web Course"

Tamara Powell
Director of Distance Education, College of Humanities and Social Sciences
Kennesaw State University
Email: tpowel25@kennesaw.edu

***Abstract*--To assist faculty at KSU (Kennesaw State University outside Atlanta, Georgia in the United States) in using instructional technology, the CHSS (College of Humanities and Social Sciences) Office of Distance Education has created, piloted, and implemented a hybrid training workshop designed to take potential online instructors from curious to comfortable and competent in three months. This workshop is offered through the KSU CHSS. This workshop design is based on secondary research into adult learning and ten years of grant supported primary research in professional development instructional technology. Five workshops have been completed so far, with the latest workshops ending January 28, 2011. Sixty two faculty in the humanities and social sciences, education, and nursing successfully completed the training. Before, during, and after the training, participants were surveyed regarding their various aspects of distance learning, including their own thoughts and beliefs. This paper presents the rationale, methods, results, and lessons learned in these trainings.**

Keywords-distance learning; hindrances to distance learning; incentives; instructional technology; motivation; professional development; training

I. INTRODUCTION

Those who direct distance learning are often called upon to share their expertise with others. We have served in various capacities where we have been called upon to assist faculty in using instructional technology. We have observed what works and have endeavored to improve upon those results. That effort has yielded some valuable insights. In 2009, the Office of Distance Education in CHSS (College of Humanities and Social Sciences) at KSU (Kennsaw State University outside Atlanta, Georgia, in the United States), created, piloted, and implemented a hybrid training workshop designed to take potential online instructors from curious about instructional technology to comfortable with instructional technology in three months, or one semester [1].

The workshop content, rationale, and research are presented here along with a link to resources for the workshop and faculty responses to each round of implementation. This workshop and its delivery methods are based on secondary research into adult learning and 10 years of grant supported primary research in training and supporting faculty in instructional technology.

A. Assessing the Need for Training

The first step in training is assessing the need for training [2]. We have found that while some faculty might do well with a handout on distance education or a book on online learning, the majority of today's faculty want a person to assist in the technological and design aspects of putting a course online. Faculty do not need in-depth training in pedagogy and assessment. They prefer training in how to transfer their successes in the physical classroom into successes in the online classroom.

B. Putting the Faculty Member in the Position of Online Student

One of the hardest things for faculty to understand is how different the orientation in the online classroom is from the orientation in the physical classroom. We all know how the physical classroom operates. The students walk into the assigned room at the assigned time and take seats. From there, the instructor tells the students how the course will progress. In the online classroom, where does the student get his or her information about the course? How does he or she know how to get started in the course or how to navigate it? How does the student communicate with the faculty member? How does he or she know how to navigate the course successfully? Having the first two sessions of this faculty development workshop meet online helps the faculty members to understand how confusing a poorly structured online course can be. This experience is intended to impress upon the faculty member the importance of setting up the course in a logical way

and letting the student know how the course should progress.

II. EXPLANATION OF THE WORKSHOP

The design of the workshop considers faculty needs and current research into andragogy to provide effective and practical training to assist faculty in the exact skills they need to transport their teaching styles into the electronic classroom.

A. Workshop Rationale: A Practical Approach

Faculty and administrators complain of high-priced "consultants" who zoom into campus to teach or explain some technological gizmo to faculty and then leave campus, leaving faculty feeling that time was wasted because 1) they didn't have enough time to explore the uses of what was taught, 2) no one applied the content to faculty needs and uses in the classroom, and 3) either unsatisfactory support or no support at all was provided after the training. In designing this workshop, our first consideration was respect for the faculty who might be involved. Faculty are adult learners, and as Malcolm Knowles established, "adults and children learn differently" [3]. Adults desire respect for their experiences, and faculty have a great deal of experience both in their subjects of expertise and in delivering education.

1) Respect for Faculty Time

Faculty are busy, and converting a course from the traditional classroom delivery method to online delivery is an added burden. Stacking an additional training workshop atop that load is unappealing to faculty. We designed this workshop to lighten the load upon faculty. For busy faculty tapped to teach online, training in instructional technology should lighten their loads considerably.

2) Respect for Faculty Expertise

Some faculty fear that distance learning will replace them with computers. During training, we emphasize the importance of individual faculty expertise in the content being developed. If faculty are tapped to teach online, no one can put their expertise online but them. Faculty can borrow and share online teaching ideas and content, but even if two faculty use the same slide presentation in their individual classes, the uses they make of it will be unique to the individual faculty member.

A breakthrough moment we had in training once occurred when we were asked to assist several faculty developing web courses who were resistant to creating content for electronic delivery. One faculty member said, "The students have a textbook, so I don't need to bother with creating my own content." The other faculty member had loaded scholarly articles and YouTube videos into his online course in lieu of creating his own content. To both, we explained, "The magical substance holding all of this course content together is your expertise. That's what students pay for, and that's what students hope to learn from. But looking at this course, we don't see your expertise. We see YouTube, we see scholarly articles by other experts, we see discussion boards where students interact with each other. All of that content is good, but one might ask of this course, what does anybody need you for?" The light bulb went off for both faculty, and their attitudes changed. After realizing how vital their decisions and expertise were to creating high-quality electronic courses, they were excited about adding their own course content, even though it is one of the most time consuming elements of creating an electronic course.

B. The Workshop

Ideally, the workshop should have been capped at 15 registered participants per session and consisted of 11 modules, running at most two hours each. However, given the high demand for this workshop, 50 faculty were originally accepted for the first run of the workshop, and 42 successfully completed the workshop. But even 42 participants, in two classes of 20-25 each, were too many.

In the evaluations of the workshop, eight participants complained that there was too wide a range of skill levels among participants. We believe that fact in itself would not have been a problem if there had not been so many participants in the workshop to begin with. Individual assistance during the workshop sessions, which is an important part of this workshop, was simply not available as it needed to be.

In the second run of the workshop, 20 faculty were accepted, and the workshop was broken into three sections of five to eight participants each.

Evaluations for the second run have not yet been completed, but faculty satisfaction seemed to be higher. Faculty still complained that there was a wide range of participant skill levels the workshop, and in response, four additional trainers are being added in the future to allow for more sections at popular time slots.

A third run of the workshop begins in January 2011 with 34 faculty participating in three sections of 10-12 faculty each. That workshop will also train three additional trainers. Two of the faculty enrolled in the third run successfully completed the first run of the workshop and are re-taking the workshop without incentives. A fourth run is planned for the fall of 2011. Data collected in every run are used to improve the next run.

In addition to four online modules and seven, two hour workshop face to face meetings, faculty are also expected and encouraged to work on their own to develop their courses in between meeting times. Faculty reported that optional, small group session "recaps" between meeting times, as well as individual sessions working one-on-one with the trainer to solve individual problems, were highly beneficial.

An example of a typical workshop is featured in the list below.

1. Online Session. Orientation to Online Learning and overview of the workshop, including streaming media lecture, discussion board, and assessment activities.
2. Online Session. Vocabulary and theory lessons. Quality Matters. Puzzles and games. Discussion board.
3. F2F (face to face) session in a computer lab: Faculty who have already taught online will share their experiences and advice and demonstrate strategies.
4. F2F Session in a computer lab: Workshop on creating content with MS Word, PowerPoint, and Adobe Professional.
5. F2F Session in a computer lab: Participants will create a web page using a free html editor (SeaMonkey).
6. F2F Session: Participants will use a Wiki (PBWiki) to critique the previous web page session. Then, participants will use the knowledge gained in the web page session to create blogs (using Blogger) and podcasts.
7. F2F Session: Participants will create streaming media (using Camtasia, Captivate, Jing, and ScreenToaster) and interactive course content (using Hot Potatoes and Quandary)
8. F2F Session: Participants will finalize their goals and assessment techniques and start to implement these items in their courses.
9. Online Session. Workshop on designing and implementing a web course, including designing banners and buttons (using Aviary).
10. Online Session. Participants will view a humorous video called "Late Night Learning with John Krutsch."
11. F2F Session: Participants will demonstrate their courses so far and discuss plans for completing their courses.

C. Requirements and Resources for Successful Implementation

The workshop requires a learning management system such as Blackboard or Moodle. KSU uses GVV (GeorgiaView/Vista). The workshop also requires a dedicated computer lab. The resources for the workshop, such as instructions and presentations, can be kept on the learning management system. However, the CHSS Department of Distance Education maintains and updates the resources on a teaching resources web page [4].

An advantage of having resources available to faculty on the open web is that faculty can access the information long after the training workshop is over or after they have moved on to other career opportunities. Faculty have told us that even a few years after taking a training workshop, they might find themselves working on a project late at night and remember a resource on the training page. They could access the resource and solve their problem instantly.

D. Marketing and Delivering a Friendly Workshop: An Open Door Policy and Rewards

Sometimes professional development trainers believe that faculty must be coerced and punished if they resist performing in the way the trainer has commanded. The real factor in such instances may not be the need for faculty compliance so much as a need on the trainer's part to feel important and

powerful. Such an attitude breeds resentment on the part of faculty.

Ultimately, blatant disrespect for faculty time and expertise is ineffective and obstructs the goal of faculty success in the electronic classroom. It is important to note that most examples of such behavior come from trainers who do not understand and respect the differences in instruction across disciplines. The trainer should deliver his or her method as a possible way to meet a desired goal rather than as the litmus test of instructor worth.

1) What Faculty Need, Not What the Trainer Wants

Faculty often balk at the idea of committing to a training workshop. Some will make the commitment because they are almost coerced by an impending online teaching assignment. Other faculty may wish to learn more about wikis, for example, and nothing else. In this training workshop, faculty are invited to drop in to any session that interests them, without committing to the entire training.

2) Rewarding Faculty

Those faculty who finish the entire workshop should be rewarded with more than a sense of achievement. The final grade rests solely on the last session of the workshop. Of course, it is not really a grade, but an incentive. To ensure faculty are on track to achieve the workshop goals, they must "show off" their progress midway through the workshop. During the sixth session on blogging, all faculty are required to post a three minute video "course so far" tour on the workshop blog [5]. Other faculty then view and comment on their colleagues' work. This project also allows others to see all the work participants are doing in the workshop.

During the last session, participants present half of the course that they have designed during the training. KSU evaluates all electronic courses using the QM rubric. Therefore, a "passing" presentation is one that provides a tour of the faculty member's course with attention to how it fulfills QM guidelines. Participants are encouraged to peer review each other's work during the presentations with their own QM rubrics. This session is a fun session, with snacks and a friendly atmosphere. Participants lavish praise on each other's work. To those who may express concern that presenting work might be humiliating to some, we would respond that a public presentation does motivate faculty to produce. If the trainer models an attitude of support throughout the workshop, then during the presentations, participants will point out the strong aspects of a colleague's work with praise before moving to suggest areas where improvement should be made.

Certainly, some faculty are more talented at design aspects, some at technological aspects, and some at pedagogical aspects. Often during the presentations, faculty find colleagues whose work inspires them, and they make plans to work together with that colleague to share resources and expertise—an outcome that we find very exciting.

Participants who attend all sessions and present part of a course at the end, including a coherent delivery plan and handout of the rest of the course plan, receive a participation certificate, QM certification, and a $3000 stipend. Also, any faculty member who attends any of the sessions receives, at a later date, a certificate listing session(s) attended.

Of course, faculty do not work for rewards such as praise, recognition, certificates or thumb drives. Also, $3000 is not enough money to compensate a faculty member for his or her time and expertise. However, despite what many students might imagine, faculty are human beings. Like most other human beings, faculty do like the idea of rewards. Sorcinelli has noted that faculty like to be recognized and rewarded and will respond positively to incentives that recognize their participation and work: "[Participants in a technology workshop for senior faculty] expressed a need for something often vaguely described as respect or recognition. Senior faculty who have been 'good citizens' and have put considerable time into developing as teachers often remark that they receive little acknowledgment for such efforts" [6]. Faculty may not realize it, but it is likely that they will receive even less acknowledgment for online efforts.

The online environment is different from the physical classroom in that it does not exist in physical space. While colleagues may see the faculty member ferrying books and teaching materials to the classroom, and hear faculty teaching, and think "Wow, what a great teacher and hard worker!" and even comment to that effect to the faculty member, online faculty will not get such positive reinforcement. In fact, the first time we taught

classes online, another faculty member remarked to us, “Oh, you’re teaching online? It must be nice to get so much time off!” Recently an administrator remarked to a group of faculty, “I just feel that faculty who teach online are getting away with something!” These comments are absurd given that creating a online course generally takes three times more time than creating a traditional course. Online faculty often complain that they feel unappreciated, so the presentation session in the training workshop may be the only time faculty get support from colleagues regarding their online work.

The Office of Distance Education in CHSS also encourages all faculty to “show off” their courses to us any time. We very much enjoy visiting with faculty and seeing their hard work and success. As a young assistant professor struggling to teach our first course online, we remember being hit hard with the realization that the administrators who assigned us this task had no idea what we were doing or how we were doing it. There were even mean-spirited whispers that we weren’t really doing anything but answering email while others were working hard at teaching in the traditional classroom. We felt very isolated, and we knew our work would never be appreciated by our peers. We want to make sure the faculty we serve do not feel that way.

D. The Importance of Post-Training Support

After the training is over and the faculty member receives his or her incentives, he or she may still need assistance with developing and implementing an online or hybrid course.

1) Encouraging the E-Faculty Community

Support for online faculty does not end when the workshop ends. The CHSS Office Of Distance Education is available on campus for personal and electronic support throughout year. Another important aspect of the training is the building of a community of faculty who teach online. Without an e-faculty community such as that which can emerge from an “e-faculty coffee hour” every month to discuss, share, and even complain about instructional technology, faculty may feel that only the trainer can assist and support him or her. Fig. 1, below, illustrates the faculty/trainer relationship that may emerge.

Figure 1. Possible model of trainer/faculty relationships after training.

Such a model does two things that should be avoided. First, it puts the trainer in a position of power, where all learning must go through him or her. While such a model might be flattering, it makes unreasonable demands upon the trainer, especially as he or she is called upon to train more and more faculty. In addition, no one person knows everything about anything. Faculty will quickly have tips and tricks to share with each other and the trainer, and such development should be fostered and encouraged. The desired model would look more like Fig. 2, below.

Figure 2. Desired model of trainer/faculty relationships after training.

2) Anytime, Anywhere Support

The method used to conduct the instructional technology workshops includes providing a hard copy handout with step-by-step instructions specific to the task that will be performed in the workshop. These instructions are created and updated by members of the KSU CHSS Distance Education Department with regard to the version of software we will be using and the order in which steps will be presented. We include, when appropriate, how to load content onto our learning management platform. In the past, we have used many different methods, including purchasing ready-made software texts for participants and putting general and specific instructions online. Nothing has worked as consistently well as creating task specific, linear instructions for faculty. Many people have suggested we try videos instead, and we did try them for some basic training in the first run. Participants commented that the videos were not helpful, and they preferred the handouts. After almost ten years of research, this method of creating goal-specific, text and graphic based, linear instructions is the one that works most effectively in assisting faculty in learning and retaining information. Research supports this observation. According to Sorcinelli, "Like most adult learners, [faculty] responded best to lots of 'hands-on' practice rather than listening to presentations" [6]. And while much visual design research supports the superiority of all graphic vs. all text instructions [7], many faculty are used to following text-based instructions and are very comfortable with them. In addition, most of the resources that we create integrate text and graphics.

When a member of the CHSS Distance Education Department is called to a faculty member's office to assist, he or she is usually greeted with the instruction sheet used in class with markings all over it. The faculty member will usually say, "I am stuck here" and point to the instruction sheet. The sheets seem to be well-used, and such specific information helps us to help the faculty member effectively.

The task specific, linear instructions are time-consuming to create, but it seems to be the best tool to help faculty save time and work more effectively. We believe that part of our office's relationship to faculty as distance learning support means that we will devote time and effort to creating resources, instituting usability tests, and updating the resources when needed. This work is one way our office shows that we respect faculty time and expertise. That respect goes a long way toward building a community of elearners.

E. Overall Perceptions of the Faculty Training Workshop

Participants were asked to evaluate their training experiences in the workshop. They were asked fifteen questions, and asked to answer "strongly agree," "agree," "disagree," "strongly disagree," or "do not know." Of the 42 participants in the first run of the workshop, 33 participated in the survey, although not all answered every question. Below, in Fig. 3-17, is a graphic summary of their responses.

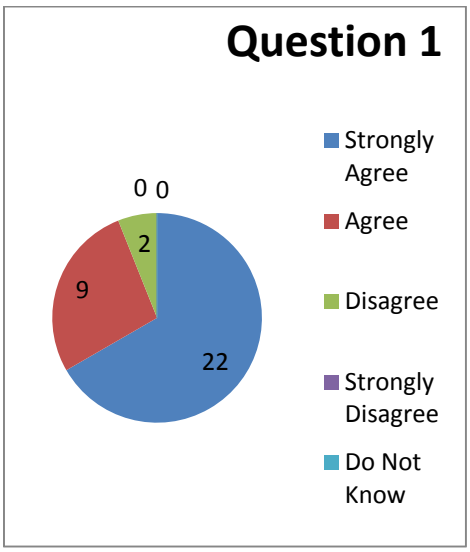


Figure 3. Question 1. The workshop provided me with useful information related to designing an online course.

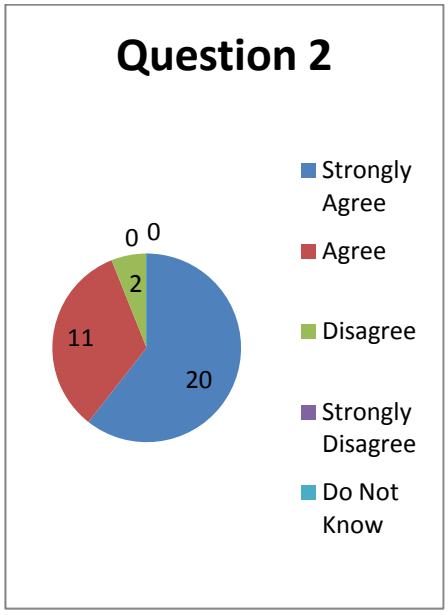


Figure 4. Question 2. The workshop provided me with useful information related to delivering an online course.

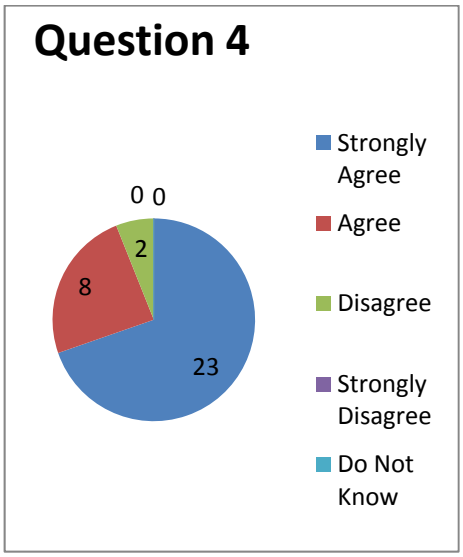


Figure 6. Question 4. The facilitator created effective components for the online portions of the workshop.

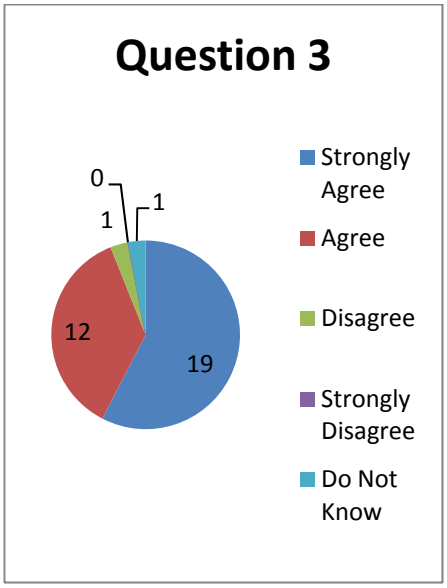


Figure 5. Question 3. The facilitator was prepared and effectively led the face-to-face portions of the workshop.

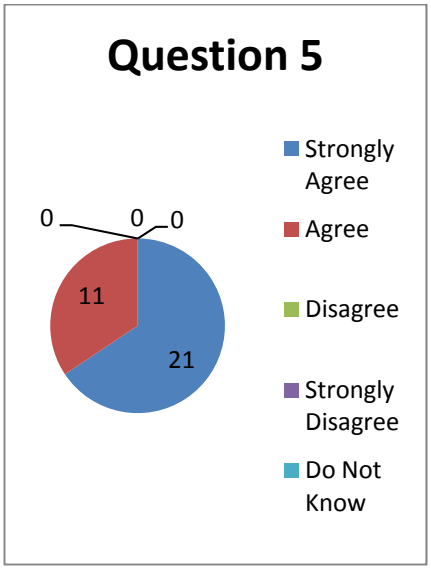


Figure 7. Question 5. The facilitator used GeorgiaView/Vista effectively in designing and delivering the workshop.

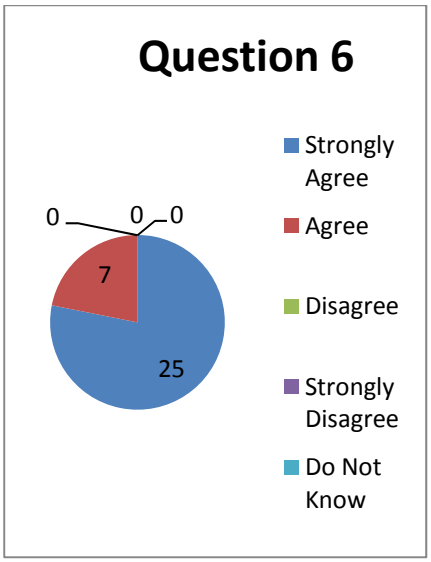


Figure 8. Question 6. The facilitator provided adequate support as I created my online or hybrid course.

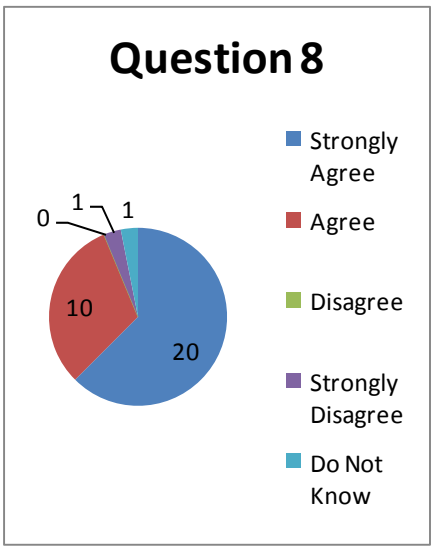


Figure 10. Question 8. The training materials provided to me during the workshop assisted me in creating course content.

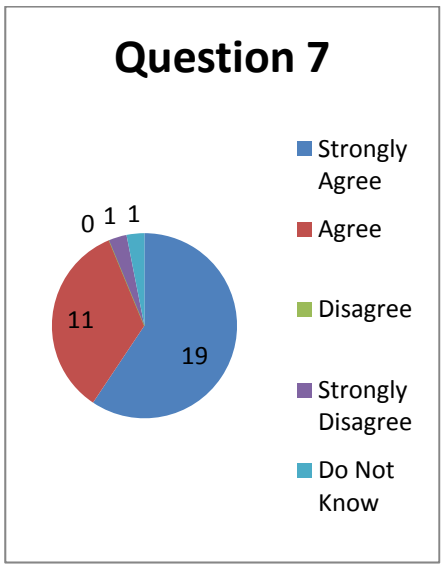


Figure 9. Question 7. The software training sessions were effective in helping me to learn what software I might choose to use in my courses.

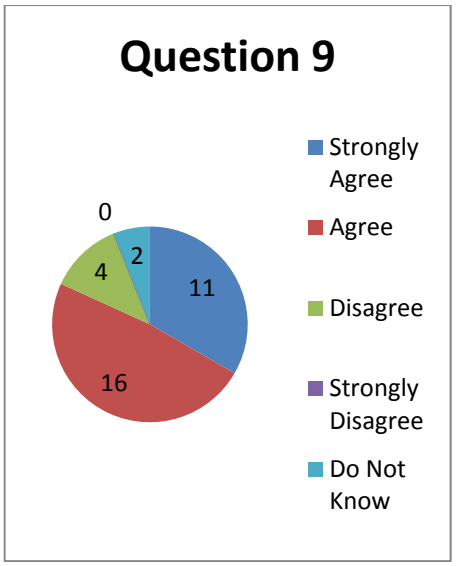


Figure 11. Question 9. The guest speakers were appropriate to the workshop and provided helpful information

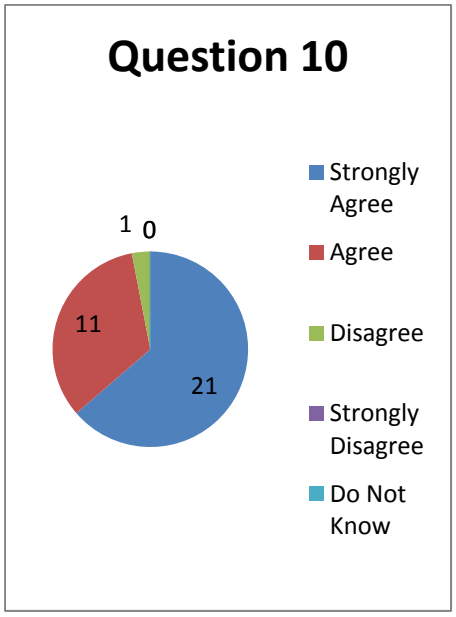


Figure 12. Question 10. My questions related to designing and delivering online courses were answered during the workshop.

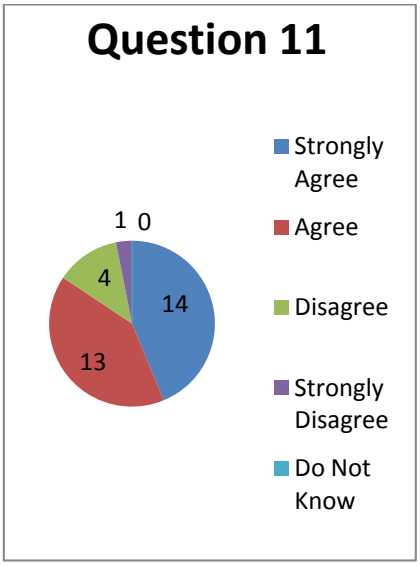


Figure 13. Question 11. The workshop prepared me to teach online.

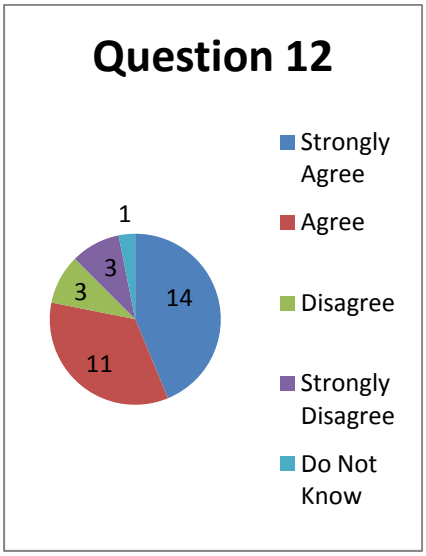


Figure 14. Question 12. The workshop prepared me to go through the QM process.

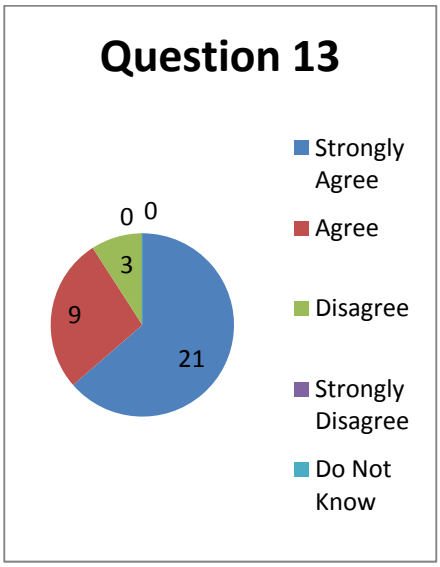


Figure 15. Question 13. If I have problems while working on my course, I know where to go or who to ask for assistance.

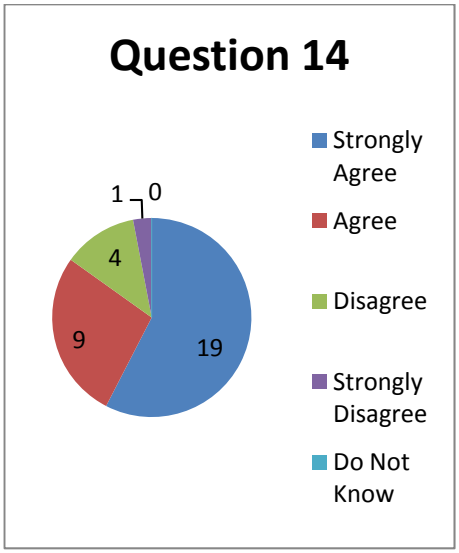


Figure 16. Question 14. Overall, I was satisfied with the workshop.

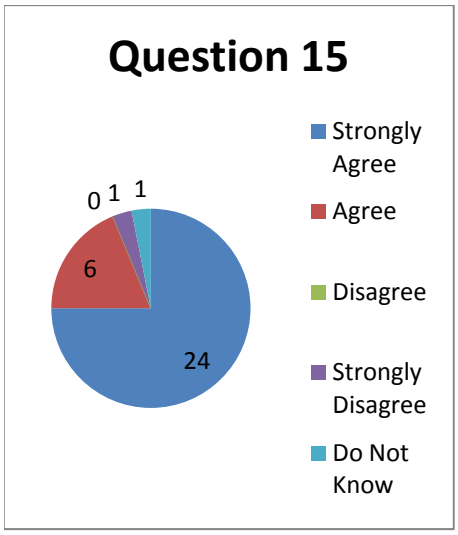


Figure 17. Question 15. Overall, I was satisfied with the facilitator.

Two additional questions were asked. Question 16 was "Identify the aspects of the workshop that most contributed to your learning (include examples of specific materials, exercises, and/or the faculty member's approach to teaching, supervision, and mentoring)." It received 31 open-ended responses, and the following summations are not necessarily taken from separate responses. Some participants mentioned many things that helped them to succeed, and some mentioned only one or two. Ten participants mentioned that the availability of the facilitator most contributed to their learning experiences. Seven participants noted that the handouts contributed the most to their learning experiences, and seven participants said the "hands on" experiences contributed the most to their learning experiences. Four participants said actually having to build components for a real course and meet deadlines for that course contributed the most to their learning experiences. One participant said that "It was helpful to have a cohort of peers who were working on the same project."

Question 17 was "Identify the aspects of the course, if any, that might be improved (include examples of specific materials, exercises, and/or the faculty member's approach to teaching, supervision, and mentoring)." Thirty participants answered the question and three skipped it, although eight answered it with "none" or "does not apply." The following summations are not necessarily taken from separate responses. Some participants mentioned many things that could be improved, and some mentioned only one or two. Nine participants mentioned that the varying skill levels of participants was a drawback, and that the class could be improved with separating participants into advanced and beginner sections. Four participants stated that more information on the QM process would improve the workshop. Three mentioned that technology problems either in their offices or in the workshops were drawbacks to their success. Two participants mentioned that one of the classrooms where one section of the workshops was held was arranged in a way that made participants choose between facing the screen or the facilitator, and that aspect was considered to be a drawback. One participant believed the workshop would have been better if it were only about GVV. Finally, three Mac users participated in the workshop, and the workshop was strongly (almost entirely) geared toward PC users. Therefore, one participant requested that Mac support be added to the workshop to improve it.

The first run of the faculty development workshop concluded in May 2010. Six months later, in November 2010, after faculty had received all their

incentives for successfully completing the workshop, faculty were again surveyed regarding their opinions of the workshop.

Of the 18 participants who responded to the survey, 66% said they used what they learned in the workshop at least once a month or more. Out of those 66%, 22% said they used what they learned more than once a week. Regarding what faculty use the most from the workshop, instructional technology information ranked first, followed by course design information and pedagogical information. One of the main goals of the workshop was to help faculty become more comfortable with teaching online. Faculty were asked,

> At the end of the workshop, participants overall rated themselves as twice as comfortable with creating blogs, wikis, websites, and audio/video materials as before they took the workshop. They also rated themselves as twice as comfortable as the control group participating in surveys. Now, seven months later, how confident do you feel about your abilities to use blogs, wikis, websites you created, and audio/video materials you created in your face to face, hybrid, and online courses?

Fifty percent responded that they were confident in their abilities. Twenty five percent rated themselves as very confident, and 6.3% said they were extremely confident ("I am the master of the Web 2.0 universe"). Only 18.8% rated themselves as "not so confident." See Fig. 18, below.

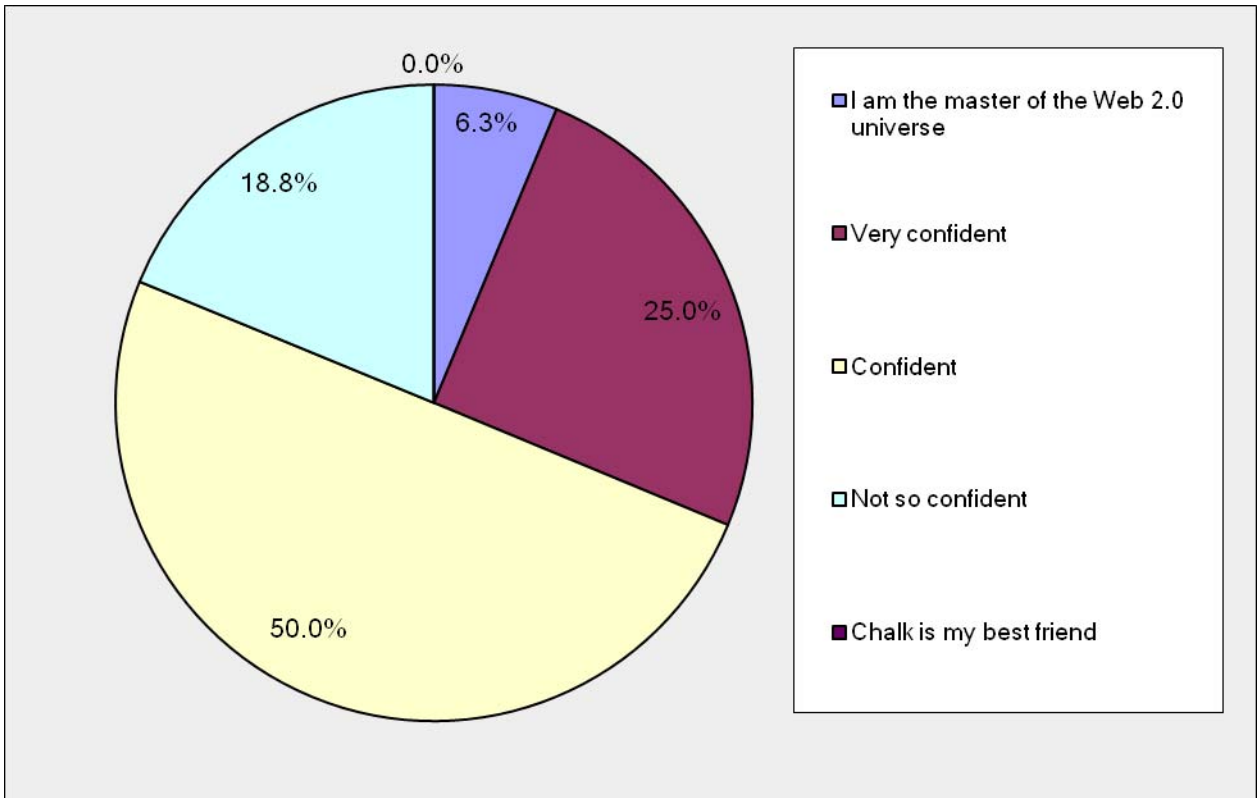


Figure 18. Six months after successfully completing the workshop, how confident do you feel about your instructional technology skills?

When asked what instructional technology tools they used the most from the workshop, faculty rated GeorgiaView/Vista first, with Camtasia second. SeaMonkey and Hot Potatoes tied for third place.

Participants were asked “On the whole, after half a year to reflect, do you feel the workshop was a valuable experience?” Ninety-three percent of responders answered “yes,” and seven percent answered “no.” When asked what could be improved, more emphasis on Mac users and products was mentioned repeatedly.

F. What Hinders Faculty from Adopting Elearning?

Many self-described “experts” on education and college faculty speculate as to why faculty do not rush to join the online course boom. What hinders faculty from embracing online learning? At conferences, we have heard these “experts” proclaim faculty to be lazy and egocentric, afraid to learn something new because they won’t know everything about it. But it turns out, the widespread public opinion that distance education is not to be taken seriously has infected the students. Therefore, some students enter online courses expecting an easy grade, and those erroneous expectations make online courses extra hard to teach. In addition, such students are often dissatisfied, and express that dissatisfaction in evaluations, which in turn jeopardizes faculty tenure and promotion.

KSU faculty who participated in this workshop were asked what hindered them from teaching online, and student attitudes topped the list. According to the KSU faculty in this survey, 31% said student attitudes was the main hindrance (see Fig. 19). The learning management system used at KSU and the poor reputation of elearning tied for second place. Other faculty attitudes came in third. All factors but the learning management system stem from stereotypes of and attitudes toward elearning, not faculty laziness and egocentrism. Only 6.3% said that mastering the technology was a hindrance. Faculty also wrote in additional responses:

1. Chair's erratic attitude. Sometimes seems supportive, then does not approve. Changes mind about whether we're allowed to offer hybrid courses.
2. Students in my hybrid commented that they hope we don't get too many hybrids and online courses at KSU because then we'll be like "them" (i.e. Univ. of Phoenix). The other hindrance is the QM...this system needs to be changed ASAP. I get tired of justifying to people outside my area (many of whom have never taught online) that I want to use "learn" "apply" or whatever verb there is. I know my field and I know how to teach. I don't want someone looking and nitpicking my course to death...especially if they don't teach online or know my field.
3. The negative attitudes of some faculty members in the English department towards online learning.
4. The huge amount of extra time in doing an online class vs an in class one.
5. The small size of our program: we need the small number of full-time faculty in our program in the traditional classroom (especially for our upper-level courses).
6. In addition to the item above, the unreasonable rigid structure of that is required for teaching online. Also, it is extremely difficult to develop a course 1 year in advance of approval.
7. Time required to put an online course together.

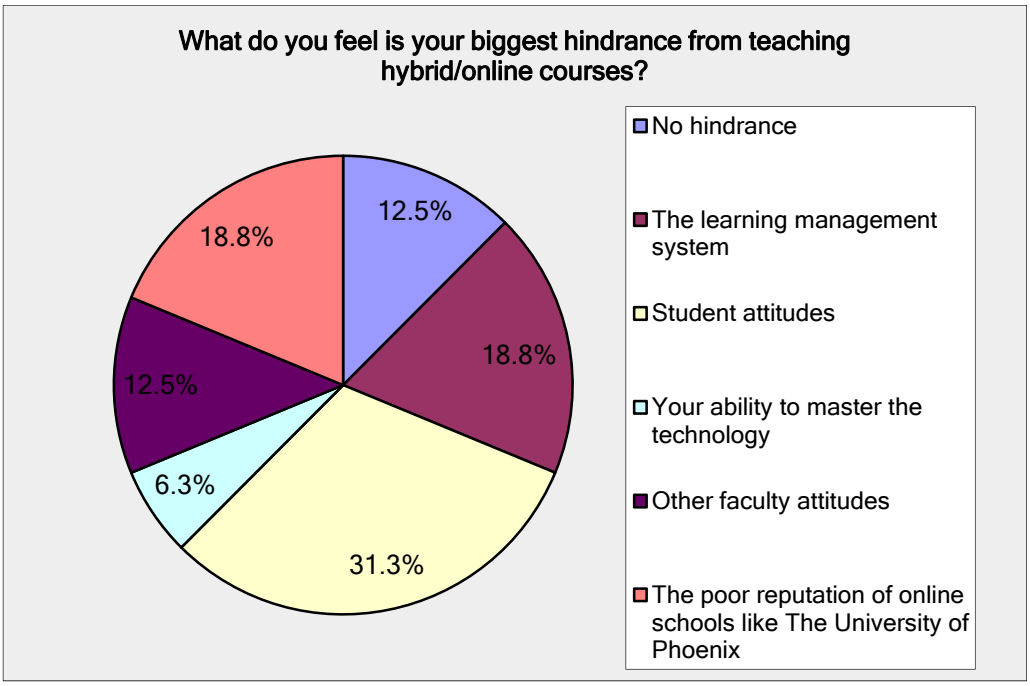


Figure 19. After successfully completing a workshop on building web courses, faculty were asked what hinders faculty from teaching hybrid/online courses. Student attitudes topped the list.

G. Does the Training Work?

The proof of success, of course, is in results. Every faculty member accepted into the workshop signs a contract stating which course will be developed into an online course or hybrid, and when it will be put taught. Therefore, 96 new online or hybrid courses should be offered in the next three years. But how many new online courses are running now at KSU because of this workshop? Of the 18 participants who responded to the survey, only half had offered the course they developed. As can be ascertained from the responses, even after developing an online course at KSU, it can be over a year before it is allowed to be offered. For that reason it is no surprise that this number is low.

Of those who had taught their courses online, 55% said they were comfortable or very comfortable teaching their courses online after the workshop.

III. CONCLUSION

We have spent the past ten years in various capacities preparing institutions to design, develop, and implement electronic learning. We have faced a multitude of scenarios, including an institution that saw distance learning as a cash cow, desired to expend as few resources as possible to develop and deliver distance education, but desired to make money hand over fist regardless. We have been hired to give a two hour workshop on distance learning to a university, only to find out that the faculty we had briefly introduced to distance learning were
expected to go out and develop programs in distance learning by the end of the next week—we quickly explained to the gentleman who contracted us that that scenario was not realistic. However, as far as he was concerned, his work, and our work, were done. The burden was now solely upon the faculty.

As many of us in the field now understand, electronic learning is an investment that must be made with full understanding that it is not a cash cow, but when developed

correctly, it can yield multi-faceted benefits. Because the burden of electronic learning falls mostly on the faculty, faculty must be a priority in distance education development. A university's desire to move into online learning is not an excuse to overburden and abuse faculty.

It is important to have institutional infrastructure and planning in place. Technological and developmental support must be available for all involved in distance learning, including faculty. Faculty also need incentives to develop, teach, and redesign online courses. Hiring a full-time team to assist faculty with instructional technology is also a good way to support faculty. Bringing in a guest speaker for an hour long presentation on "what is distance learning" is fine, but it is not a solid leg of support for faculty who are expected to create online courses and programs. And certainly money to bring in the guest speaker would be better spent as incentives for faculty to complete instructional technology training.

All faculty use faculty evaluations to improve their courses, and this workshop is no different. After the first run of the workshop, faculty feedback resulted in 1) more support for Mac users, 2) fewer participants in each workshop session, 3) more self-check and quiz items on areas that participants seemed to overlook, 4) "what's next" components to help participants orient themselves at the end of each module, and 5) small weekly prizes to help participants stay motivated and engaged.

In the faculty evaluations of the first workshop, although the entire second online module of the workshop addressed QM, three participants commented that QM was not addressed early in the workshop. One participant also added to his/her comments that "I do not think that the power point [sic] presentations with voice were effective. They took quite a while to listen to and much of the information could have been presented in a quicker format that would have been easier to review at a later date." This comment stands in stark contrast to all other comments about that content delivery mode. This comment also seemed to indicate that the participant did not listen to the online modules and, therefore, missed the QM information presented early in the workshop. Subsequent runs of the workshop have included recaps of online modules in the face to face meetings to encourage participants to take the online workshop materials as seriously as the face to face activities. Faculty who teach hybrid courses often remark that students "dismiss" the online portions of hybrid courses as "not important." Clearly, some faculty attitudes parallel those of our students.

Formal evaluation data from the second run of the workshop is not yet available. However, informal participant feedback resulted in 1) more trainers being added to each workshop session, 2) printable activity checklists for each online module, 3) use of the calendar function in GVV to help participants gauge what tasks should be completed each week, 4) the addition of the Hybrid-O-Matic, a online tool developed by the CHSS Department of Distance Education that counsels faculty regarding how to translate their face to face teaching styles into a hybrid course, 5) an advanced workshop series focusing on those with a higher level of skills, and 6) brown bag workshops led by faculty experts in various techniques and technologies, mainly focusing on hybrid learning. We expect that improvements will be made for each run for the life of the workshop.

To help address the negative attitude toward elearning both from the public, from faculty, and from students, KSU has launched a distance learning website to support students and faculty interested in online learning [8]. A video, "Words of Wisdom from KSU Online Faculty," located prominently on the site, works hard to dispel the myth that online courses are not serious courses.

As a result of these training workshops, we had hoped to see more of a building of an elearning community, but invitations to socialize outside of class were met with replies of "We're too busy," as indeed, faculty are. Faculty did socialize to a degree on the workshop wiki at the beginning of the workshop, and perhaps more electronic social networking opportunities will better serve faculty needs. That area is ripe for future research.

In short, there's no easy, fast, and cheap solution to moving faculty toward creating quality online instruction. However, the investment and time are worth it, as online education becomes part of every university's offerings. Especially in the United States, a lack of affordable child care, coupled with the rising cost of health care, including elder care, in part, drives the demand for distance education opportunities. Many adults want educational opportunities but can't leave children and elderly parents alone and travel to the university. For the increasing number of adults in such situations, distance education is a necessity. The demand will continue to increase.

The workshop described in this paper won the 2010 Sloan Consortium Award for Faculty Development in Online Teaching.

ACKNOWLEDGMENT

Great thanks to KSU CHSS, and Dean Richard Vengroff and Associate Dean Thierry Leger especially, for sponsoring and supporting these workshops and this research. Thank you to Dr. Solomon Negash for encouraging us in this research. Thank you to KSU's CETL (Center for Excellence in Teaching and Learning), ITS (Information Technology Services), ODG (Online Development Group), and DLC (Distance Learning Center) for being the backbone of distance learning at KSU. Thank you to all members of the KSU CHSS Distance Education team: Dr. Laura McGrath, Ms. Nikki Hill, and Mr. Dustin Proctor. Also, great thanks to the Sloan Consortium, who recognized our hard work. Finally, thanks to all of the faculty who have supported us as trainee and trainer in the past ten years. We appreciate your patience.

REFERENCES

[1] T. Powell, "Delivering Effective Faculty Training: A Course and Methods to Prepare Faculty to Teach Online," Proc. elml 2010 Second International Conference on Mobile, Hybrid, and On-Line Learning, Feb. 2010, pp.7- 10, doi: http://doi.ieeecomputersociety.org/10.1109/eLmL.2010.8. Accessed February 20, 2011.

[2] R. Hassell-Corbiell, Developing Training Courses: A Technical Writer's Guide to Instructional Design and Development. Tacoma, WA: Learning Edge, 2001.

[3] M.S. Knowles, E.F. Holton III, and R.A. Swanson, The Adult Learner. 6th ed. Burlington, MA: Elsevier, 2005.

[4] T. Powell. "Kennesaw State University College of Humanities and Social Sciences Electronic TeachingResources.http://www.kennesaw.edu/elearning.Accessed February 20, 2011.

[5] "Impressive Elearning Course Tours." http://24hourstours.blogspot.com. Accessed February 20, 2011.

[6] M.D. Sorcinelli, "Post-tenure Review Through Post-tenure Development: What Linking Senior Faculty and Technology Taught Us." Innovative Higher Education vol. 24.1, Fall 1999, pp. 61-72. Academic Search Complete. EBSCO. Horace W. Sturgis Library, Kennesaw, GA. 25 Aug. 2009.

[7] E. Donaldson, L. Shoemaker, J. Slavin, and J. Lake. "Text Vs. Graphics Formats for Online Instructions." Student HCI Shore'99 Online Research Experiments. University of Maryland. Department of Computer Sciences. August 28, 2009. http://otal.umd.edu/SHORE99/jdfl/index.html. Accessed January 15, 2011.

[8] "Distance Learning." Kennesaw State University. 2010. http://www.kennesaw.edu/distancelearning.shtml.

Last accessed February 20, 2011.

.

A Process Model for Establishment of Knowledge-Based Online Control of Enterprise Processes in Manufacturing

Daniel Metz, Sachin Karadgi, and Manfred Grauer
Information Systems Institute,
University of Siegen,
Siegen, Germany
Email: {metz, karadgi, grauer}@fb5.uni-siegen.de

Abstract **- Today's enterprises are operating in a complex and volatile business environment. To address this situation, enterprises endeavor to realize horizontal and vertical integration of enterprise processes (i.e., business and manufacturing processes) leading to a real-time enterprise. Research has been carried out in integrating data generated from automation devices in (milli) seconds and transactional data from business applications, referring to long time horizons (e.g., days). However, this integrated data is not used extensively for online control of enterprise processes. Therefore to overcome this issue, a process model is presented for identification and assimilation of knowledge. This process model comprises four steps: (i) analysis and (re-) design of enterprise processes to be controlled, (ii) creation of enterprise data model and data flow diagrams for automation devices and business applications, (iii) (offline) knowledge identification based on knowledge discovery in databases process, and (iv) online monitoring and control of enterprise processes using complex event processing. The envisaged process model is a prerequisite for implementation of IT-framework used for online monitoring and control of enterprise processes. Both, the process model and the corresponding IT-framework have been implemented and validated in a casting enterprise.**

Keywords - enterprise integration, knowledge discovery in databases, data mining, online control, complex event processing.

I. INTRODUCTION

The business environment of an enterprise has become complex, volatile and mainly driven by uncertainties [1]. In addition, pressure on an enterprise to manufacture components with high quality, reduced lead time and low cost has been intensified. As a consequence, relevant enterprise (value creation) processes (i.e., business and manufacturing processes) have to be flexible, adaptable and controlled online. In this regard, the integration of enterprise processes in horizontal and vertical direction of an enterprise has become indispensable. Available enterprise application integration (EAI) systems can be used to horizontally integrate existing business applications like enterprise resource planning (ERP) systems, supply chain management (SCM) systems, and customer relationship management (CRM) systems [2]. Along with this horizontal integration, vertical integration of different enterprise levels can be seen as a prerequisite for establishing online control of enterprise processes [3], and the vision of a real-time enterprise (RTE) [4][5].

According to German standard VDI 5600 [6], an enterprise can be classified into different manufacturing execution system (MES) levels as illustrated in Figure 1: (i) enterprise control level, (ii) manufacturing control level, and (iii) manufacturing level. VDI 5600 focuses on the problems and benefits related to MES. Similarly, standards like IEC 62264 [7] are available and emphasis on realization of MES. In the current contribution, various terminologies are based on VDI 5600.

At enterprise control level, business processes are performed to achieve the enterprise's long term strategies. Thus, business processes can be designed, configured, enacted, and analyzed applying four steps of a business process management (BPM) life cycle: (i) business process design, (ii) business process configuration, (iii) business process enactment, and (iv) business process diagnosis [8][9]. Process-aware information systems (PAIS) like workflow management systems (WMS) are in charge to invoke business applications (e.g., ERP system) and (web) services along a workflow execution (i.e., automation of a business process) to fulfill certain enterprise's strategic objectives [8][9]. During business process enactment at enterprise control level, planned performance values (i.e., TO-BE values) are generated offline (i.e., in months or weeks) and further, these values are transactional [10].

On the contrary, manufacturing processes are employed to accomplish the objectives set at the enterprise control

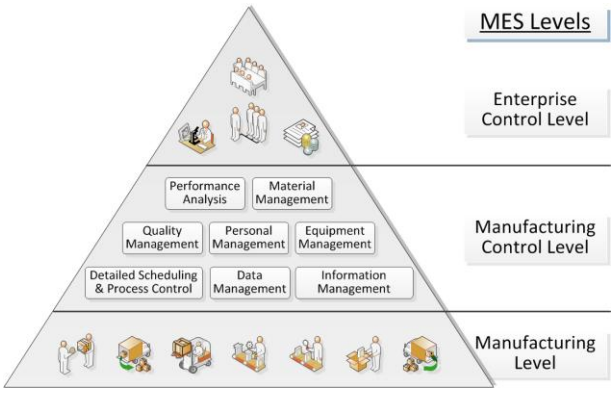


Figure 1. Enterprise levels as defined in VDI 5600 [6].

level. Automation devices are available at manufacturing level to execute manufacturing processes. Enormous amount of data (e.g., sensor data) is generated by these devices during execution of manufacturing processes in real-time (i.e., seconds or milliseconds). In addition, operators provide necessary data related to automation devices or orders like pre-defined reasons for a resource breakdown, order details during start of order execution, and so forth. Overall, these values (i.e., AS-IS values) indicate the actual performance at the manufacturing level.

MES solutions and business process applications are available to achieve computerized and automated vertical integration at certain MES levels [11]. However, major problems still remain open with respect to the interface between the enterprise control level and the manufacturing level [12]. More precisely, the realization of enterprise-wide multi-loop control within and across all levels from enterprise control level to manufacturing level is not adequately achieved [10][12]. Also, inadequate vertical integration hinders the establishment of enterprise-wide knowledge and learning cycles [13]. First, data from different MES levels is not adequately integrated and thus, it cannot be exploited to identify new knowledge. Second, if any new knowledge has been derived, it is not incorporated in online control of enterprise processes. As a consequence, concepts of RTE: sense-and-respond and learn-and-adapt are integrated insufficiently into enterprise processes [14].

The current contribution is based on the IT-framework for digital enterprise integration [13], and presents a methodology for identification and assimilation of knowledge for online control of enterprise processes in manufacturing. State-of-the-art on enterprise integration (EI), monitoring and control of enterprise processes, and data mining in manufacturing is summarized in Section II. A novel methodology is elaborated in Section III to establish enterprise-wide knowledge and learning cycles for online control of enterprise processes. Section IV describes an industrial case study. Finally, conclusion and future work are discussed in Section V.

II. STATE-OF-THE-ART

The methodology elaborated in Section III is based on various concepts like EI and data mining. Therefore, the current section summarizes state-of-the-art research in the area of EI, enterprise data model, identification of knowledge, and utilization of knowledge for online monitoring and control of enterprise processes.

Around the mid 1990's, several EI reference architectures (e.g., CIMOSA, PERA, ARIS and GRAI/GIM) were available to guide the design and implementation of an integrated enterprise [15]. However, these reference architectures were different in terms of their theoretical background [15], and enterprise understanding, modeling approaches, and purposes [16]. Hence, GERAM - generalized enterprise reference architecture and methodology has been developed to address these differences [17]. Aforementioned reference architectures have contributed in defining GERAM and later, it was standardized as ISO 15704 - requirements for enterprise reference architecture and methodologies [18].

The reference architectures mentioned above do not reveal how to realize them in terms of technologies. Apart from enterprise reference architectures, several software vendors have developed MES solutions to bridge the vertical integration gap between MES levels, like MES-HYDRA [11]. But also with MES, the exchange of data between MES levels is done manually or at most semi-automatically due to inflexible and proprietary interfaces [12][19].

An agent-based production monitoring and control system (PMC) based on the JADE framework was elaborated [20]. The PMC Provis.Agent integrates various IT-systems and machine control devices, and establishes the use of information between various systems (e.g., for visualization). Also, NIIIP-SMART architecture provides horizontal and vertical integration, and interoperation utilizing workflow, enterprise rules, agents and STEP [21].

Service orientation (especially by means of web and grid service technology and their corresponding standards) has been used for EAI [22]. In service oriented architecture (SOA), business applications offer their functionalities as services. These services can be loosely coupled and orchestrated to complex workflows. Because of the loosely coupled structure, the realized IT-architecture is flexible and adaptable. Hence, European-funded projects SIRENA [23] and SOCRADES [24] aim to exploit this SOA paradigm to seamlessly integrate heterogeneous resources located at manufacturing level with business applications at enterprise control level. In this regard, a prototype for vertical integration of SOA-ready devices with SAP MII was presented [5]. Unlike the predominant request-reply communication approach of traditional SOA, an enterprise has to react to events online, and hence, necessitates implementation of publish-subscribe mechanism [25]. However, this doesn't make SOA obsolete as SOA and event processing are complementary concepts for achieving modularity, loose-coupling, and flexibility [26].

Enterprise data model is necessary to enable EI, i.e. to relate TO-BE and AS-IS values from different MES levels. IEC 62264-2 [7] describes models and terminologies that enable to implement enterprise control between enterprise control level and manufacturing control level. Further, this standard can be augmented with various technical models like DIN EN 61512-2 [27]. An MES database structure for system integration was analyzed with respect to resource, system plans, system status and system configurations [28]. A factory data model was defined to represent strategic intent, capability, organization structure and behavior of an enterprise [29]. This factory data model consists of strategy, facility, process, resource, token and flow classes. The token classes represent physical flow of material (e.g., work piece, invoice). Flow classes represent links to token classes and process classes.

Knowledge is embedded into enterprise processes by enterprise members (i.e., know-who) in form of know-what, know-why, and know-how [30]. However, this knowledge is tacit and context-specific. This knowledge is externalized by enterprise members and can be expressed by means of

enterprise process data (i.e., TO-BE values). These process data are updated online during execution of enterprise processes in terms of feedbacks (i.e., AS IS values). TO-BE and AS-IS values from different MES levels are integrated for online control of enterprise processes [13][31]. These integrated values are stored in a relational database as historical data, which is periodically exploited in offline processes to calculate key performance indicators (KPIs) and overall equipment effectiveness (OEE), among others.

Extensive research has been carried out to convert tacit knowledge embedded in processes into explicit knowledge using analytical methods (e.g., data mining), but this research focuses mostly on transactional data (e.g., finance, sales) at the enterprise control level (e.g., [32]). Around 7% of data mining methods are utilized to address problems in manufacturing [33]. This limited usage of data mining in manufacturing enterprises originated in the perception of relatively high efforts to achieve EI [33]. Nonetheless, data mining methods have been used in manufacturing domains like manufacturing system, and maintenance [34][35].

A decision making process related to enterprise process control can be a complex task spread across different MES levels, and depends upon the quantity and quality of information. In this regard, a framework for organizing and applying knowledge for decision making in manufacturing and service applications was elaborated [36]. The decision making process was supported with the knowledge derived using data mining algorithms.

European-funded project K-NET (sub-project of Future Internet Enterprise Systems (FInES) cluster) has presented an approach at a conceptual level to enhance, monitor and reuse of knowledge in a networked enterprise [37]. In addition, an enterprise modeling and integration framework was presented based on knowledge discovery in databases (KDD) by extending the views of CIMOSA i.e., adding knowledge and mining views [38]. In most of the enterprises, explicit knowledge is codified as rules managed in rule-based systems (RBS) [39].

RBS like Drools Expert [40] are often lacking in taking temporal and causal relations between events into account. As a consequence, research attempts are been made to employ the externalized knowledge for creating event patterns utilized in a complex event processing (CEP) engine. Unified event management architecture was conceptualized to deal with primitive and complex events for monitoring and control of manufacturing processes [41]. Nonetheless, the architecture is positioned at manufacturing control level and integrates real-time data from manufacturing level but neglects to integrate transactional data from enterprise control level. Further, architecture for an extensible event-driven manufacturing system was elaborated [42]. This architecture was built on a MES solution with a tight integration with enterprise control level and manufacturing level, and utilized CEP engine to manage events triggered in manufacturing level. However, the presented approach does not address knowledge identification required to define event patterns.

System Insights provides an open source framework for online monitoring and analysis of manufacturing enterprises [43]. The framework constitutes following components - data delivery, data collection, and data analysis. Data delivery from different devices is achieved through MTConnect standard and MTConnect data bus. Data is stored in (high speed) databases using functionality of data collection. For control of enterprise processes, data analysis is performed online utilizing the services of EsperTech [44] or Drools Fusion [45] CEP engines. Also, the stored data is utilized offline to calculate various metrics.

III. METHODOLOGY

An IT-framework for digital enterprise integration was articulated [13][31]. This framework utilizes tracking objects along with a RBS. However to consider the temporal and causal relations between events triggered across various MES levels, it is essential to replace RBS with state-of-the-art CEP engine. On basis of this framework, a methodology

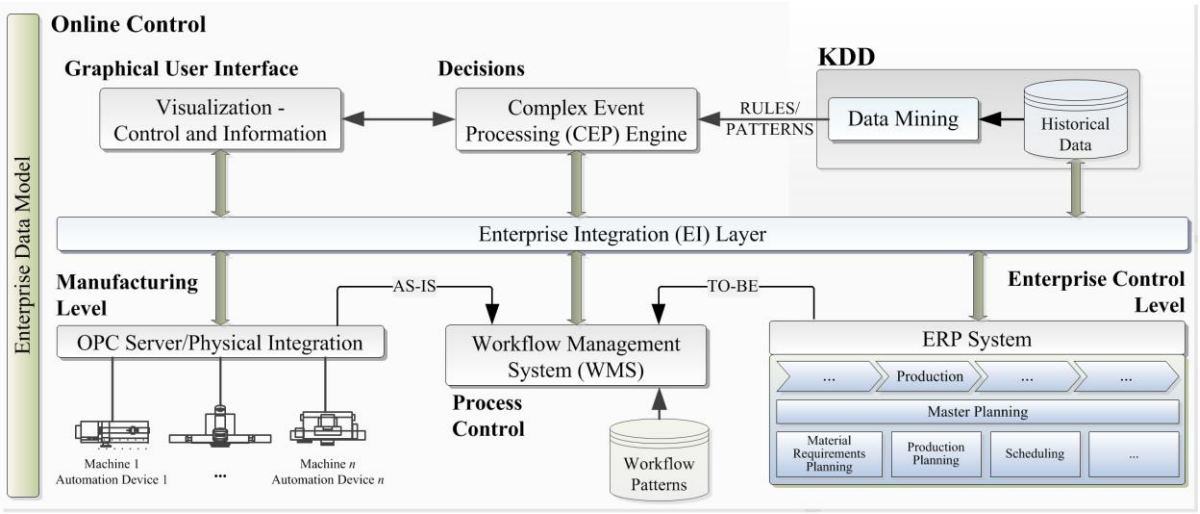


Figure 2. IT-framework for digital enterprise integration (adapted from [13]).

is elaborated for identification and assimilation of knowledge for online control of enterprise processes. This methodology encompasses following components: (i) enterprise process analysis, (ii) enterprise data model and data flow diagram, (iii) knowledge identification, and (iv) assimilation of knowledge for online control of enterprise processes. In the following subsections, the aforesaid IT-framework is introduced in brief and further, the components of the methodology are elaborated.

A. *Overview of IT-Framework for Digital Enterprise Integration*

An architectural overview of the IT-framework for digital enterprise integration is illustrated in Figure 2. In addition, flow of process data and control data between various components of this IT-framework is depicted in Figure 3. The IT-framework is based on available standards (e.g., ISO 15704 [18], IEC 62264 [7]), technologies and paradigms (e.g., SOA paradigm) and involves various IT-systems (e.g., ERP system). Enterprise processes are instantiated from predefined workflow patterns (see Step 1 in Figure 3) and are supplied with necessary planned performance or process values (i.e., TO-BE values) from business applications like ERP systems (see Figure 2 and Step 2 in Figure 3). Business applications or at least their crucial functionality in a certain context (e.g., accessing planned performance values) are made available as services within an SOA.

As one of the purposes of the IT-framework is online control of enterprise processes, publish-subscribe mechanism usually applied in event-driven architectures (EDA) is implemented. EI layer subscribes to the events triggered by manufacturing resources (i.e., automation devices) at the manufacturing level. Three-tier architecture for physical integration has been implemented to collect data from these manufacturing resources and forward the collected data to all the subscribers, here EI layer (see Step 3 in Figure 3 and [13]). The received real-time data from the manufacturing level denotes actual achieved performance (i.e., AS-IS values). Planned performance values from enterprise control level (e.g., ERP system) together with actual achieved performance values from manufacturing level are integrated according to an enterprise data model, and stored in relational database as historical data (see Step 6 in Figure 3).

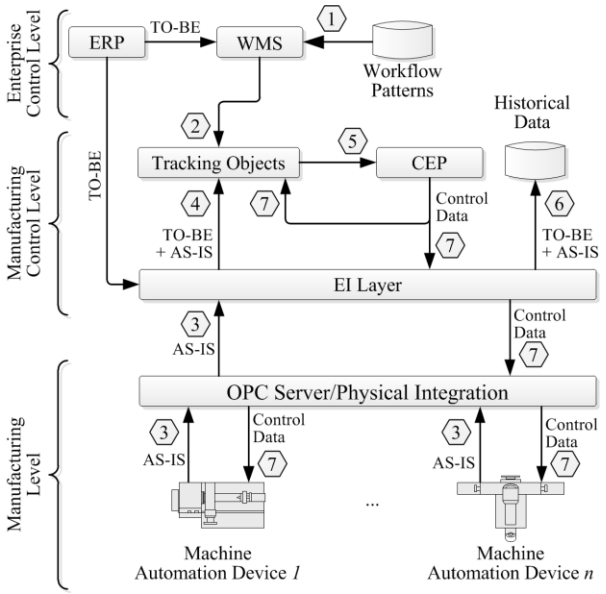


Figure 3. Flow of process data and control data in IT-framework for digital enterprise integration.

The digital enterprise integration framework was further enhanced for online monitoring and control of enterprise processes using tracking objects [31]. Tracking objects represent control-relevant objects like orders, products and resources, and are instantiated simultaneously with a corresponding workflow instance in a WMS (see Step 2 in Figure 3). These objects are updated during enterprise process execution with the values acquired from different MES levels (see Step 4 in Figure 3). The changes in tracking objects' status can be analyzed online by a RBS (e.g., Drools Expert) [31]. However, the usage of RBS implicates the lack of taking temporal and causal relations between events into account. In addition, it is remarkable that only a few WMS support the collection and interpretation of real-time data [8][9]. Hence, the usage of a CEP engine (e.g., EsperTech [44]) instead of RBS is necessary. Here, the CEP engine is in charge of continuously analyzing tracking objects and dispatching control data to required MES levels (see Step 7 in Figure 3).

In addition to the control of enterprise processes employing CEP engine, the historical data is periodically utilized offline to calculate KPIs and OEE. However, historical data is seldom exploited to identify new knowledge and further this identified knowledge is not utilized for online monitoring and control of enterprise processes. To overcome the aforesaid drawbacks, a methodology is elaborated to identify and assimilate knowledge for online monitoring and control of enterprise processes.

B. *Methodology for Knowledge Identification and Assimilation in Manufacturing*

An overview of the process model towards the realization of digital enterprise integration has been depicted in Figure 4. This process model consists of four process steps, which are unique to a particular enterprise. The process model can be put into practice by means of implementing the aforementioned IT-framework. These process steps need not necessarily be performed sequentially and further, individual process steps can be carried out from time to time to enhance enterprise (value creation) processes.

Prior to the implementation of the IT-framework for digital enterprise integration (see Section III.A), it is essential to analyze and (re-) design the enterprise processes as it is in the case of BPM life cycle [8][9] (see Step I in Figure 4). An enterprise data model based on industrial standards (e.g., IEC 62264 [7]) and enlarged with technical models (e.g., DIN EN 61512-2 [27]) is in charge of relating AS-IS and TO-BE values from different MES levels (see Step II in Figure 4). Also, data flow diagrams (DFDs) can be created to reveal the interdependencies between business

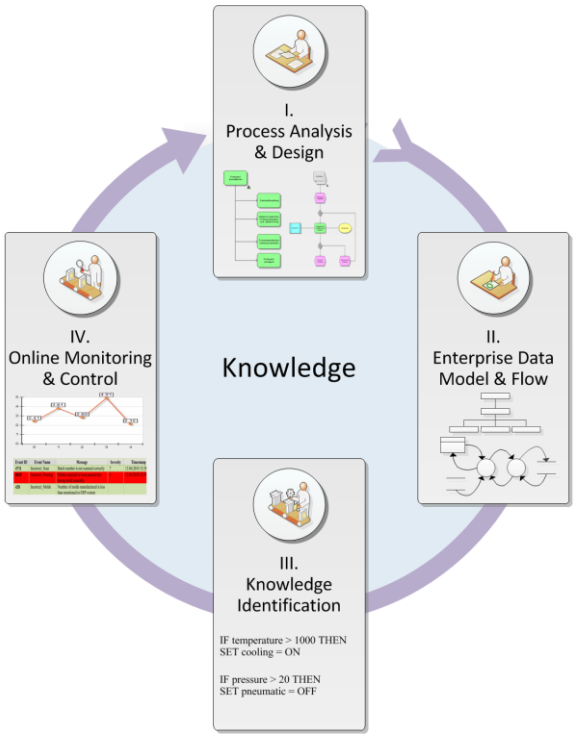


Figure 4. Process model towards the realization of knowledge-based online control of enterprise processes.

applications and resources, and associated events triggered at various MES levels.

Knowledge is embedded in enterprise process data (e.g., pressure, temperature) generated before and during execution of these processes i.e., TO-BE and AS-IS values. These process data is mapped onto the aforementioned enterprise data model and stored in a relational database as historical data. Subsequently, offline KDD process can be employed on the historical data to identify new knowledge (see Step III in Figure 4). Applying repeated and time-consuming database queries executed on the integrated database are not useful for online control of enterprise processes [46]. Instead, event streams (i.e., TO-BE and AS-IS values) created during process execution need to be analyzed and processed online using a CEP engine (see Step IV in Figure 4). Here, the externalized knowledge can be codified as event patterns and event pattern rules, and these event patterns can be used for detection of complex events in event streams. Along with this detection of complex events, event processing can be employed for online control of enterprise processes. In following subsections, each of the process steps will be elaborated.

1) Enterprise Process Analysis and (Re-) Design

Enterprise process analysis and (re-) design is part of enterprise reference architectures (e.g., ARIS [47]) and business process reengineering (BPR) [48]. BPR can be described using four main phases: (i) identification of critical enterprise processes, (ii) review, update and analysis of enterprise processes (AS-IS analysis), (iii) (re-) design of enterprise processes based on AS-IS analysis, and (iv) implementation of (re-) designed enterprise processes.

Comprehension of enterprise processes and their integration within an enterprise's organizational structure is crucial for the implementation of online enterprise process control strategies. Hence, process analysis and (re-) design incorporates enterprise's organizational structure as well as its process-oriented organization.

The activities and functions of an enterprise process are executed by various resources (e.g., IT-systems), which are in charge of an enterprise's organizational unit. Thus, the organizational units can be organized and modeled using organizational charts. In addition, functions and activities of an enterprise process take several kinds of inputs. These can be data (e.g., printed documents) but also intangible inputs like implicit knowledge of enterprise members. In summary, each function of an enterprise process can be linked with an organizational unit, various inputs and outputs. Further, the functions are orchestrated to enterprise processes using logical connectors like 'and', 'or', and 'exclusive or'.

Several modeling languages and methodologies are available to model enterprise processes like event-based process chain (EPC) and business process modeling language (BPML). Apart from these modeling languages that focus on tangible inputs and outputs (e.g., data and documents), knowledge management description language (KMDL) can be employed to describe knowledge intensive processes, i.e., creation, use and necessity of knowledge along enterprise processes [49].

2) Enterprise Data Model and Data Flow Diagram

Today's business applications and automation devices are complex, and several inputs are required by these systems to define enterprise processes. In addition, enormous amount of data is generated by the automation devices in real-time denoting information like feedbacks, product positions and alerts. For online monitoring and control of these enterprise processes, it is essential to analyze the business applications and automation devices, and their corresponding processes to identify critical control-related process parameters. In this regard, enterprise data modeling is an essential step to establish important control-relevant parameters. It influences the quality of information that is necessary to execute enterprise processes, achieve EI, and enhance online monitoring and control of enterprise processes.

IEC 62264-2 describes models and terminologies that enable to implement enterprise-wide control between enterprise control level and manufacturing control level [7]. An enterprise data model based on IEC 62264-2 can be further augmented with technical models depending on the type of manufacturing process, like DIN EN 61512-2 (for batch manufacturing, [27]) and DIN 8582 (for metal forming processes, [50]). Overall, IEC 62264-2 facilitates to exchange structured data between business applications and manufacturing resources.

Besides the static structure of an enterprise data model, DFDs reveal the interdependencies between processes' systems and manufacturing resources, either in isolation or in

combination [51]. Further, DFDs identify the flow of data, describing the dynamic behavior of enterprise processes. Process data is generated and manipulated by several resources during process execution. The DFDs can be employed to expose the relationships between various IT-systems and resources. As the number of systems and resources has been increased in industrial scenarios, DFDs are usually organized in a hierarchy of DFDs. Coarse-grained DFD can depict an overview of the shop floor and its resources while a certain DFD is been created with regard to a certain enterprise (sub-) process.

3) Knowledge Identification

Knowledge can be defined from different perspectives. In the current research context, following definition is adapted: "Data is raw numbers and facts, information is processed data, and knowledge is authenticated information" [39]. As mentioned earlier, knowledge is embedded into enterprise processes as process data. This knowledge is enriched and enlarged during execution of enterprise processes (e.g., feedbacks). Further, integrated enterprise process data (i.e., TO-BE and AS-IS values) is stored as historical data based on the aforementioned enterprise data model.

Historical data can be also exploited to derive new knowledge, which can be utilized in online monitoring and control of enterprise processes. Hence, knowledge identification can be performed utilizing KDD process, defined as a "process of mapping low-level data into other forms that might be more compact or abstract or useful" [52]. KDD process is depicted in Figure 5. Input to KDD process is historical data and outputs are patterns, subjected to certain defined quality known as interestingness measures [53]. These interestingness measures can be objective measures based on the statistical strengths or properties of the discovered patterns and subjective measures, which are derived from the user's beliefs or expectations [53]. A pattern is an abstract representation of a subset of data and needs to be evaluated by domain experts to identify knowledge. KDD process consists of several steps (see Figure 5), which are elaborated in the following paragraphs.

Understanding of manufacturing domain in concern is indispensable for successful employment of KDD process. This can be carried out as described in process analysis and (re-) design, and enterprise data model and DFDs (see Step I and II in Figure 4). Major activities of manufacturing

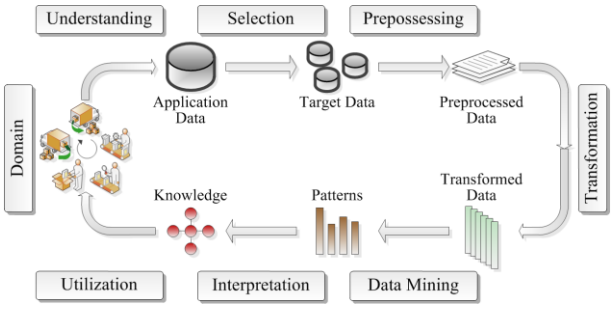


Figure 5. Knowledge discovery in databases (KDD) process (adapted from [51]).

enterprises are production, maintenance, quality and inventory [7], and consequently, define the goals of the KDD process. Depending upon the goals of the KDD process, specific data (target data) is selected from the historical data on which patterns will be searched.

However, historical data might be inaccurate due to various reasons and thus influencing the identified knowledge. During data acquisition process, data is collected from operators through console, and analog equipments and digital measuring devices through programming logic controllers (PLCs) (see Figure 2). In the aforementioned IT-framework, AS-IS values are made available to enterprise integration layer through object linking and embedding (OLE) for process control (OPC) severs (see Figure 2 and Step 3 in Figure 3). Data collected consists of noise or inaccuracies or missing values, which makes searching of patterns complicated [54]. This might be due to limitations of measuring instruments, typing errors of operator or errors in logic of PLCs. To overcome these inaccuracies, it is necessary to study and understand the domain, as described in Step I in Figure 4. Understanding of domain along with the enterprise data model will help to identify suitable statistical methods to remove noise, strategy to fill the missing values and deletion of duplicate data. Also periodically, data collection can be enhanced by verifying the collected data from manufacturing resources in coordination with the operators. Overall, cleaning and pre-processing activities are carried out on the selected data for further processing.

Only subset of pre-processed data is required for achieving the aforementioned goals of KDD process [55]. Hence, transformation of data involves reducing number of parameters in the target data or representing the target data in a more general or acceptable format. Filter and wrapper approaches can be employed to reduce number of parameters [56]. In addition, understanding of manufacturing processes, operations, and constraints will aid in transformation process supported with enterprise data model and DFDs.

Data mining is a particular sub-process in KDD process. It is based on proven techniques like machine learning, pattern recognition, statistics, artificial intelligence, knowledge acquisition, data visualization, and high performance computing [57]. Data mining consists of three steps: (i) selection of data mining method, (ii) determination of appropriate data mining algorithms, and (iii) employing these algorithms for pattern search.

Data mining methods have to be chosen in accordance to the goals of KDD process as they determine the type of knowledge to be mined i.e., concept description, classification, association, clustering and prediction [34]. Since the goal in current research is to enhance online monitoring and control of enterprise processes, knowledge to support decision making processes needs to be identified. In this regard, classification and regression methods can be engaged to determine new knowledge. Classification is a method to categorize a new instance of data into one of several predefined classes [52][54]. It consists of two steps [34]: (i) construction of a (classification) model based on the analysis of database tuples (i.e., a training set) described by

attributes and (ii) usage of this constructed (classification) model for classification of new data instances. In contrary, regression is a function that maps data to a real-valued prediction variable.

According to previously selected data mining methods (e.g., classification), data mining algorithms have to be selected to search for patterns. For example, decision trees, decision rules, inductive logic programming and rough set methods can be utilized to determine rules [36]. Finally, the previously selected data mining methods and algorithms are employed to discover patterns.

Discovered patterns can be sufficient large or it might be necessary to select a subset of discovered patterns [56]. Consequently, to enhance the quality of discovered patterns, measure of interestingness can include both subjective and objective approaches [56]. A discovered pattern can be interpreted using structured interviews with domain experts. If necessary, all or selected steps of KDD process need to be repeated in order to obtain more suitable knowledge. If sufficient integrated data is not available to carry out KDD process, structured interviews with domain experts can be conducted to identify initial knowledge. Later, acquired historical data can be exploited employing the aforementioned KDD process to enhance and enlarge the knowledge base.

4) Knowledge-Based Online Control of Enterprise Processes

Historical data can be employed to identify new knowledge as described in Section III.B.3. This identified knowledge can be used for control of enterprise processes by directly accessing the aforementioned historical data. However, repeated and time-consuming database queries [46] (e.g., applying ex-post online analytical processing (OLAP) queries) result in offline control of enterprise processes. Consequently, real-time data should be processed online (i.e., near real-time) for control of enterprise processes utilizing previously identified knowledge, tracking objects and CEP engine as depicted in Figure 2 and Figure 3.

Tracking objects are representatives of process entities (e.g., products, orders, resources) of a particular enterprise process. A tracking object is instantiated simultaneously with a workflow in a WMS specifying a process route (see Step 1 in Figure 3) and associated planned performance values (i.e., TO-BE values). During execution of enterprise processes, AS-IS values from manufacturing level are made available in OPC servers and simultaneously forwarded to EI layer utilizing publish-subscribe mechanism (see Step 3 in Figure 3). These AS-IS values are integrated with corresponding TO-BE values from ERP system obtained using request-reply mechanism.

EI layer manages the integrated process data simultaneously in numerous ways. First, tracking objects are updated with corresponding integrated process data (see Step 4 in Figure 3) and thereby, tracking objects contain up-to-date status information of an actual enterprise entity within an enterprise process. Second, integrated process data is delivered to all subscribed clients with graphical user interface (GUI) for online monitoring of enterprise processes (see Figure 2). Finally, integrated process data is stored in a relational database as historical data for offline analysis (see Figure 2 and Step 6 in Figure 3). Tracking objects are constantly analyzed at manufacturing control level and utilized for online control of enterprise processes using CEP engine with the objective to enhance major activities of manufacturing enterprises i.e., production, maintenance, quality and inventory (see Step. 5 in Figure 3). Subsequently, CEP engine is in charge of dispatching control data to manufacturing resources (see Step 7 in Figure 3), and at the same time updating tracking objects with control data. In the following paragraphs, CEP and assimilation of previously identified knowledge is elaborated.

An event is characterized by its event source (e.g., a certain automation device), event type, event attribute (i.e., data) and timestamp or time interval [58] and additionally, event sink (e.g., operator, plant manager) as depicted in Figure 6. As mentioned previously, events are triggered across different MES levels during execution of enterprise processes and form an event cloud [43][58]. Events can be classified as simple or composite events based upon their level of abstraction. Simple events are triggered across different MES levels and do not have any abstraction. Hence, a simple event does not provide sufficient information for online control of enterprise processes [41]. For instance, a simple event is triggered whenever a lower mold is produced by a molding machine.

On contrary, a composite event with higher abstraction can be described with an event pattern based on simple events [41][59]. For example, a complex event is triggered whenever total of number of molds produced for a given order exceeds the required quantity specified in the order. Further, higher abstraction events can be derived from composite events as depicted in Figure 7. In summary, a composite event can be "created by combining base events using a specific set of event constructors such as disjunction, conjunction, sequence, etc" [58]. Finally, an event stream is a "linearly ordered sequence of events", which are ordered by arrival time or bounded by a certain time interval [58]. Here, event streams are composed of simple events denoting tracking objects created or updated during process execution.

An event pattern is a "template containing event templates, relational operators and variables" [58]. Relationships between different (simple and composite) events can be basic, temporal and spatial [43] or logical, temporal and causal [41][60]. Simple events are created at a

```
<event>
  <event_id>1234</event_id>
  <event_name>MOLD NUMBER</event_name>
  <event_type>MACHINE</event_type>
  <event_time>12:23:45.11</event_time>
  <event_data>234309</event_data>
  <event_causality>-</event_causality>
  <event_source>MOLDING 1</event_source>
  <event_sink>PLANT MANAGER</event_sink>
</event>
```

Figure 6. Simple event description in an XML format.

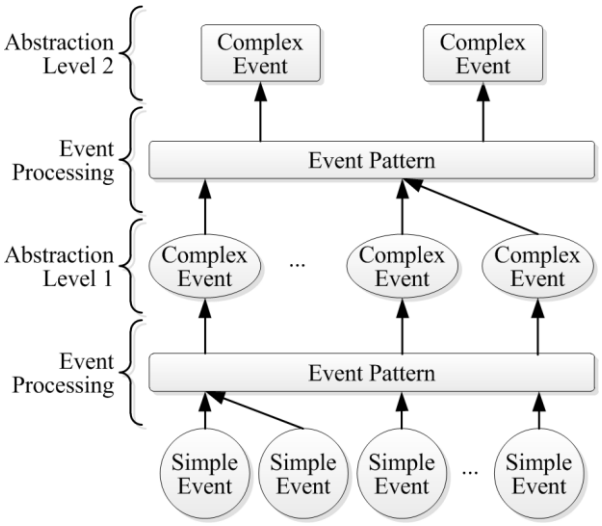


Figure 7. Hierarchical event abstraction (adapted from [58]).

certain period in time i.e., have an associated event timestamp. However, enterprise members from enterprise control level and manufacturing control level are interested in aggregated events for online monitoring and control of products, orders, or resources [41][60]. To enable aggregation of events, it is necessary to utilize temporal event patterns and hence, support time interval with sliding time boundary [41][60], as depicted in Figure 8. Temporal event patterns include overlap, coincides, contains and before or after event patterns [43]. Further, operators (e.g., concatenation, sequence) associated with temporal event patterns can be identified [41].

CEP can be defined as "computing that performs operations on complex events, including reading, creating, transforming or abstracting them" [58][60]. The main purpose of the CEP engine is to control enterprise processes based on a continuous analysis of events streams (i.e., tracking objects). As described before, tracking objects contain up-to-date status information of an enterprise process entity. On update with the new incoming integrated data, tracking objects are analyzed within the CEP engine as shown in Step 5 in Figure 3. Event patterns expressed by an event processing language (EPL) are used within the CEP engine, which is capable to analyze logical, temporal, and

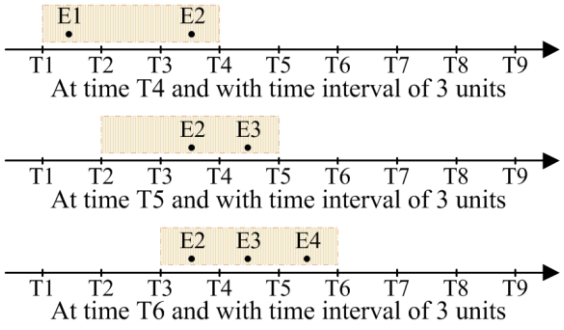


Figure 8. Processing of temporal events using sliding time boundary.

causal event patterns. Further, event pattern rules (i.e., reactive rules) define how the CEP engine reacts to the occurrence of a certain event pattern [59]. An example of an event pattern and event pattern rule is shown in Table I, based on [59]. Hence, the incoming events are analyzed using event pattern rules and necessary control data is dispatched. In addition, suitable events with higher abstractions are created and appended to the already existing event cloud for future event processing.

TABLE I. EVENT PATTERN AND EVENT PATTERN RULE

Element	Declarations
Variables	Order O, Order.Quantity Q, MoldList LM, MoldList.Count C, Mold M, Mold.Order M_O
Event types	MoldProduced(Order O, Mold M) Produce(Order O)
Pattern	MoldProduced (O, M)
Context test	C < Q
Action	Create Produce(O)

In order to effectively monitor and control the enterprise processes, it is essential to identify and characterize events. Previously identified knowledge can be assimilated by creating event patterns and event pattern rules codified as EPL statements. In addition, structured interviews with the domain experts can be utilized to enhance and enlarge event patterns and corresponding event pattern rules. Finally, event patterns and event pattern rules can be made persistent in a centralized database. Enterprise members or decision makers are not interested in all the events. Hence, event sinks or event consumers can be configured by enterprise members' roles (e.g., supervisor, plant manager) and their corresponding privileges (e.g., defined in a lightweight directory access protocol (LDAP) server). Therefore, an event configuration, part of client's GUI provides the necessary functionality to define and configure events and event patterns.

There are two implementations on how the CEP engine influences or controls the actual enterprise processes. First, the CEP engine uses interfaces and services provided by the EI layer (see Figure 2) to automatically dispatch control commands. Second, before manipulating enterprise processes, CEP engine exposes envisaged decision as a suggestion to clients with GUI. Here, an enterprise member accepts or declines the proposition. Obviously, human interaction is used in cases where enterprise members should take liability. However, access to the aforementioned functionality of dispatching control data depends upon the enterprise members' roles and their corresponding privileges.

IV. INDUSTRIAL CASE STUDY

The IT-framework as well as the corresponding process model for enabling digital enterprise integration and achieving online control of enterprise processes elaborated in Section III can be put into practice in different types of manufacturing, especially in batch manufacturing (e.g.,

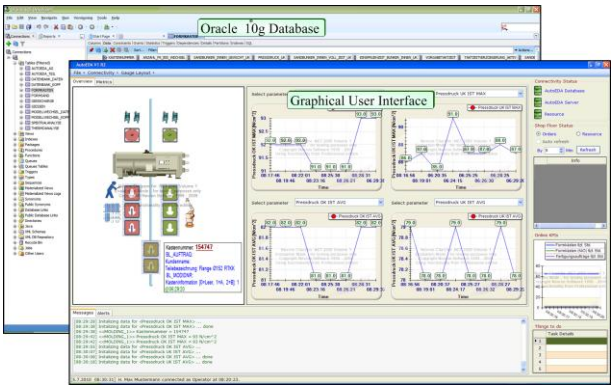


Figure 9. Software screenshots of implemented IT-framework for enabling enterprise integration and control of enterprise processes.

casting processes) and discrete manufacturing (e.g., sheet metal forming processes). Here, an attempt is made to realize the framework for casting processes with special purpose resources. To efficiently utilize these capital-intensive resources, online monitoring and control of enterprise processes is mandatory. The IT-framework has been implemented using Microsoft™ Visual Studio IDE and .NET framework 3.5. Different screenshots of the implemented IT-framework, stacked one over other, are displayed in Figure 9.

Enterprise processes have been analyzed and modeled using ARIS (utilizing EPC and Entity-Relationship-Model) [47]. IEC 62264-2 [7] and DIN 61512-2 [27] have been adapted to create an enterprise data model. In addition, DFDs were created to reveal interdependencies, and dynamic behavior between various automation devices and business applications.

Data is acquired from different automation devices and made available as OPC items in OPC servers (see Figure 2). This data is forwarded to EI layer. Here, the data is mapped onto the enterprise data model and integrated with TO-BE values from ERP system. EI layer manages the integrated data in numerous ways. First, integrated data is delivered to all clients with GUI for online monitoring of enterprise processes. The subscription of clients to process data is realized through a windows communication foundation (WCF) interface. Delivered data is displayed online by the clients using visual elements like charts and gauges. Second, integrated values are stored in an Oracle® 10g database for offline process analysis (see Figure 10). Client's GUI provides interfaces to track products, orders and resources using request-reply mechanism (i.e., accessing historical data). Finally, tracking objects containing integrated data are processed in EsperTech CEP engine [44] for online control of enterprise processes, especially with the objective to enhance productivity.

Casting process consists of following sub-processes: molding to manufacture molds, melting of raw material (e.g., aluminum), pouring of molten material, cooling and finishing (e.g., cutting, grinding, reaming). Molding machine in consideration is capable of producing upper and lower molds at high production rate and influences upstream as

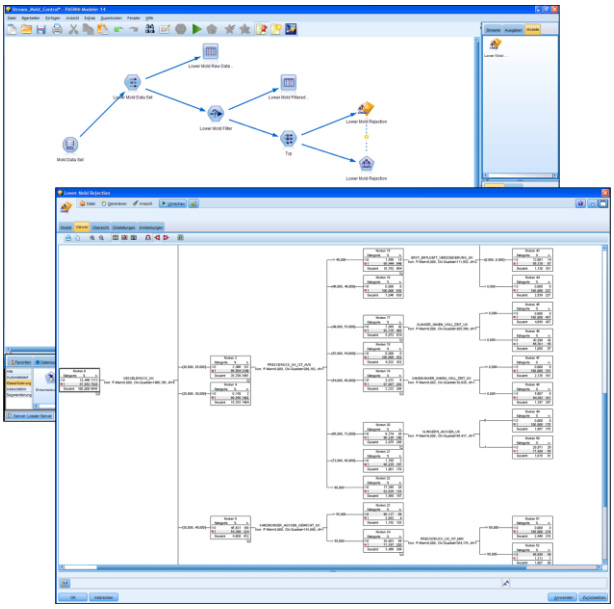


Figure 10. IBM® SPSS® Modeler Professional screenshots.

well as downstream processes. Therefore, goal of the KDD process is to enhance productivity of molding process by reducing number of rejects. Thus, enhancing productivity and reducing wastage (i.e., sand, molten material). Numerous applications are available to support steps of KDD process as illustrated in Figure 5. Here, IBM® SPSS® Modeler Professional [61] has been chosen.

At start of KDD process, a suitable data set has been selected from Oracle® 10g database for utilizing in SPSS® Modeler as illustrated in Figure 10. Here, data set containing mold details is selected, with 14891 rows. Further, each mold detail is associated with 459 attributes, stored in different database tables. Molding machine simultaneously produces lower and upper molds, and here, lower mold details are considered for further analysis. Hence, structured interviews with domain experts were carried out to identify the attributes of lower mold. Therefore, only 39 attributes of lower mold are retained and extraneous attributes are discarded in SPSS® Modeler. Still, reduced data set might contain erroneous data, unexpected situations or many unrelated parameters. SPSS® Modeler provides different graphical tools (e.g., histogram, distribution) to analyze and identify aforementioned causes from the data set. Also, filtering expressions are utilized to clean the data set. After these preprocessing and transformation, data set contains 12799 rows.

Classification algorithms like Chi-square Automatic Interaction Detectors (CHAID) algorithm provided by SPSS® Modeler are used to construct decision trees. Suitable target attribute (e.g., mold quality – good or bad) is selected. Remaining 38 attributes are chosen as input parameters. After executing the CHAID algorithm, a decision tree is created with depth of decision tree equal to 4 and 12 predictors (e.g., pressure). These determined patterns contained in decision tree are validated and refined by domain experts, i.e., using subjective interestingness

```
insert into AlarmEvent
select ProductionRessourceName,
ProductionProcessDataName, ProductionProcessData,
'Pressure is too high' as AlarmMessage from
ProcessDataEvent(ProductionRessourceName='MOLDING_1'
and ProductionProcessDataName='Pressure').win:length(2)
having avg(cast(ProductionProcessData, double)) >
cast(22.5, double)
```

Figure 11. Event pattern codified as an EPL statement in EsperTech.

measure. Further, validated patterns are utilized in creating event patterns and event pattern rules codified in EPL statements as shown in Figure 11. For instance, creation of an alarm event by analyzing applied mold pressure on last two molding machine's simple events. These EPL statements can be managed by domain experts with suitable privileges via a client's GUI and stored in an EPL database.

V. CONCLUSION AND FUTURE WORK

Today's enterprise environment is complex, volatile and driven by uncertainties, forcing enterprises to become more flexible and adaptable. Consequently, enterprises endeavor to overcome the aforesaid challenges by enhancing the online monitoring and control of their enterprise processes. This can be achieved by integrating the enterprise along horizontal and vertical direction. As a consequence, transactional data from enterprise control level and real-time data from manufacturing level have to be integrated and stored as historical data in relational database. However, these historical data is not used extensively for identification of knowledge and subsequently, identified knowledge is not used in the control of enterprise processes.

In the current contribution, a process model for identifying and assimilating knowledge for online control of enterprise processes has been presented. This process model consists of four components: (i) enterprise process analysis, (ii) enterprise data model and DFDs, (iii) knowledge identification, and (iv) assimilation of identified knowledge for online control of enterprise processes. These process steps need not necessarily be performed sequentially and individual process steps can be carried out from time to time to enhance enterprise value creation processes.

Enterprise processes are analyzed and modeled following the ARIS approach. Available standards were adapted to derive the enterprise data model. Analyzed enterprise processes assisted in creating DFDs revealing interdependencies between various automation devices and business applications. Real-time data from manufacturing level and transactional data from enterprise control level are integrated based on enterprise data model and stored in an Oracle® 10g database. Also, integrated data was displayed to enterprise members using charts and gauges. Knowledge was identified using IBM® SPSS® Modeler Professional. Further, the identified knowledge was assimilated for online control of enterprise processes using EsperTech CEP engine.

Currently, the framework has been used in an enterprise for online monitoring and control of batch manufacturing (i.e., casting processes). Future implementation is planned for discrete manufacturing processes i.e., for an automotive sheet metal component supplier.

ACKNOWLEDGMENT

Parts of the work presented here have been supported by German Federal Ministry of Economics and Technology (BMWi) as part of "Central Innovation Programme SME" (ZIM) initiative (KF2111502LL0). Also, we are thankful to our industrial partner Ohm & Häner Metallwerk GmbH & Co. KG, Germany for the opportunity to implement the elaborated methodology and framework in a casting enterprise. Especially, we would like to acknowledge Dr.-Ing. Ludger Ohm, Dr.-Ing. Georg Dieckhues, and Jürgen Alfes for their valuable comments and support.

REFERENCES

[1] M. Grauer, D. Metz, S. S. Karadgi, and W. Schäfer, "Identification and Assimilation of Knowledge for Real-Time Control of Enterprise Processes in Manufacturing," Proc. 2nd Int'l Conf. on Information, Process and Knowledge Management (eKNOW 2010), Feb. 2010, pp. 13-16, doi: 10.1109/eKNOW.2010.14.

[2] D. Linthicum, Enterprise Application Integration, Addison-Wesley Longman, Amsterdam, 2000.

[3] J. Lee, K. Siau, and S. Hong, "Enterprise Integration with ERP and EAI," Comm. of the ACM, vol. 46, no. 2, Feb. 2003, pp. 54-60.

[4] A. Drobik, M. Raskino, D. Flint, T. Austin, N. MacDonald, and K. McGee, The Gartner Definition of Real-Time Enterprise, tech. report, Gartner Inc., 2002.

[5] S. Karnouskos, D. Guinard, D. Savio, P. Spiess, O. Baecker, V. Trifa, and L. de Souza, "Towards the Real-Time Enterprise: Service-based Integration of Heterogeneous SOA-ready Industrial Devices with Enterprise Applications," Proc. 13th IFAC Symp. on Information Control Problems in Manufacturing (INCOM '09), June 2009, pp. 2127-2132.

[6] VDI 5600, Manufacturing Execution System (MES) - VDI 5600 Part 1, Verein Deutscher Ingenieure (VDI), 2007.

[7] IEC 62264, Enterprise-control system integration, All Parts.

[8] W. M. P. van der Aalst, A. H. M. ter Hofstede, and M. Weske, "Business Process Management: A Survey," Proc. 1st Int'l Conf. on Business Process Management, Springer-Verlag, Berlin, LNCS 2678, 2003, pp. 1-12.

[9] M. Weske, W. M. P. van der Aalst, and H. M. W. Verbeek, "Advances in Business Process Management," Data & Knowledge Eng., vol. 50, no. 1, July 2004, pp. 1-8, doi:10.1016/j.datak.2004.01.001.

[10] A. Kjaer, "The Integration of Business and Production Processes," IEEE Control Systems Magazine, vol. 23, no. 6, 2003, pp. 50-58.

[11] J. Kletti, Ed., Manufacturing Execution System – MES, Springer, Berlin, 2007.

[12] H. Panetto and A. Molina, "Enterprise Integration and Interoperability in Manufacturing Systems: Trends and Issues," Computers in Industry, vol. 59, no. 7, Sept. 2008, pp. 641-646, doi: 10.1016/j.compind.2007.12.010.

[13] M. Grauer, D. Metz, S. Karadgi, W. Schäfer, and J. W. Reichwald, "Towards an IT-Framework for Digital Enterprise Integration", Proc. 6th Int'l Conf. on Digital Enterprise Technology (DET 2009), AISC, vol. 66, Springer, Berlin, Dec. 2009, pp. 1467-1482, doi: 10.1007/978-3-642-10430-5_111.

[14] C. Meyer, "Keeping Pace with Accelerating Enterprise", CIO Insight, Nov. 2002; http://www.cioinsight.com/c/a/Expert-Voices/Expert-Voice-Christopher-Meyer-on-the-Accelerating-Enterprise/, 25.12.2010.

[15] A. Chen and F. Vernadat, "Standards on Enterprise Integration and Engineering - State of the Art," Int'l J. of Computer Integrated

Manufacturing, vol. 17, no. 3, Apr. 2004, pp. 235-253, doi: 10.1080/09511920310001607087.

[16] S. Aier, C. Riege, and R. Winter, "Enterprise Architecture - Literature Overview and Current Practices," Wirtschaftsinformatik, vol. 50, no. 4, 2008, pp. 292-304 (in German).

[17] P. Bernus and L. Nemes, "The Contribution of the Generalised Enterprise Reference Architecture to Consensus in the Area of Enterprise Integration," Proc. of ICEIMT97, K. Kosanke and J. Nell, Eds., Springer, 1997, pp. 175-189.

[18] ISO 15704, Requirements for Enterprise Reference Architecture and Methodologies, ISO 15704:2000/Amd 1:2005, 2005.

[19] S. Karnouskos, O. Baecker, L. de Souza, and P. Spiess, "Integration of SOA-ready Networked Embedded Devices in Enterprise Systems via a Cross-layered Web Service Infrastructure," Proc. 12th IEEE Int'l Conf. on Emerging Technology and Factory Automation, Sept. 2007, pp. 293-300.

[20] O. Sauer and G. Sutschet, "Agent-Based Control", Computing & Control Eng. J., vol. 17, no. 3, June 2006, pp. 32-37, doi:10.1049/cce:20060305.

[21] C. Gilman, M. Aparicio, J. Barry, T. Durniak, H. Lam, and R. Ramnath, "Integration of Design and Manufacturing in a Virtual Enterprise Using Enterprise Rules, Intelligent Agents, STEP and Workflow," Proc SPIE Int'l Symp. on Intelligent Systems & Advanced Manufacturing , 1997, pp. 160-171.

[22] D. Linthicum, Next Generation Application Integration: From Simple Information to Web Services, Addison-Wesley Professional, Amsterdam, 2003.

[23] H. Bohn, A. Bobek, and F. Golatowski, "SIRENA - Service Infrastructure for Real-Time Embedded Networked Devices: A Service Oriented Framework for Different Domains," Proc. Mobile Comm. and Learning Technologies, Apr. 2006.

[24] L. M. S. de Souza, P. Spiess, D. Guinard, M. Köhler, S. Karnouskos, and D. Savio, "SOCRADES: A Web Service Based Shop Floor Integration Infrastructure," Internet of Things, C. Floerkemeier, M. Langheinrich, E. Fleisch, F. Mattern, and S. Sarma, Eds., Springer, Berlin, LNCS 4952, 2008, pp. 50-67.

[25] R. Schulte, A Real-Time Enterprise is Event-Driven, tech. report, Gartner Inc., 2002.

[26] S. T. Yuan and M. R. Lu, "A Value-Centric Event Driven Model and Architecture: A Case Study of Adaptive Complement of SOA for Distributed Care Service Delivery", Expert Systems with Applications, vol. 36, no. 2, Mar. 2009, 3671-3694.

[27] DIN EN 61512-2, Batch Control - Part 2: Data Structures and Guidelines for Languages, Ref. Nr. DIN EN 61512-2:2003-10, 2003.

[28] B. Zhou, S. Wang, and L. Xi, "Data Model Design for Manufacturing Execution System," J. of Manufacturing Technology Management, vol. 16, no. 8, 2005, pp 909-935.

[29] J. A. Harding, B. Yu, and K. Popplewell, "Information Modelling: An Integration of Views of a Manufacturing Enterprise," Int'l J. of Production Research, vol. 37, no. 12, Aug. 1999, pp 2777-2782.

[30] W. M. Cheung and P. G. Maropoulos, "A Novel Knowledge Management Methodology to Support Collaborative Product Development," Digital Enterprise Technology - Perspectives and Future Challenges, P. F. Cunha and P. G. Maropoulos, Eds., Springer, June 2007, pp. 201-208.

[31] M. Grauer, S. S. Karadgi, D. Metz, and W. Schäfer, "An Approach for Real-Time Control of Enterprise Processes in Manufacturing using a Rule-Based System," Proc. Multikonferenz Wirtschaftsinformatik, Feb. 2010, pp. 1511-1522.

[32] C. Groba, I. Braun, T. Springer, and M. Wollschlaeger, "A Service Oriented Approach for Increasing Flexibility in Manufacturing," Proc. 7th IEEE Int'l Workshop on Factory Communication Systems, Communication in Automation (WFCS 2008), Dresden, May 2008, pp. 415-422.

[33] H. Kuntze, T. Bernard, G. Bonn, and C. Frey, "Entscheidungsunterstützung im Produktionsumfeld mit Data-Mining-Werkzeugen," VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA): AUTOMATION 2008 Lösungen für die Zukunft, June 2008 (in German).

[34] A. Choudhary, J. Harding, and M. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge," J. of Intelligent Manufacturing, vol. 20, no. 5, Oct. 2009, pp. 501-521, doi: 10.1007/s10845-008-0145-x.

[35] J. Harding, M. Shahbaz, Srinivas, and A Kusiak, "Data Mining in Manufacturing: A Review," J. of Manufacturing Science and Engineering, vol. 128, no. 4, Nov. 2006, pp. 969-976, doi:10.1115/1.2194554.

[36] A. Kusiak, "Data Mining: Manufacturing and Service Applications," Int'l J. of Production Ressearch, vol. 44, nos. 18-19, Sept./Oct. 2006, pp. 4175-4191, doi:10.1080/00207540600632216.

[37] E. Mazharsolook, S. Scholze, S. Ziplies, R. Neves-Silva, and K. Ning, "Enhancing Networked Enterprise Management of Knowledge and Social Interactions," J. of Computing in Systems and Eng., vol. 10, no. 4, 2009, pp. 176-184.

[38] E. Neaga and J. Harding, "An Enterprise Modelling and Integration Framework Based on Knowledge Discovery and Data Mining," Int'l J. of Production Research, vol. 43, no. 6, Mar. 2005, pp. 1089-1108, doi: 10.1080/00207540412331322939.

[39] M. Alaavi and D. E. Leidner, "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues," MIS Quarterly, vol. 25, no. 1, Mar. 2005, pp. 107-136.

[40] Drools Expert, http://www.jboss.org/drools/, 25.12.2010.

[41] K. Walzer, J. Rode, D. Wünsch, and M. Groch, "Event-Driven Manufacturing: Unified management of Primitive and Complex Events for Manufacturing and Control", IEEE Int'l Workshop on Factory Communication Systems, 2008.

[42] Y. H. Zhang, Q. Y. Dai, and R. Y. Zhong, "An Extensible Event-Driven Manufacturing Management with Complex Event Processing Approach", Int'l J. of Control and Automation, vol. 2, no. 3, Sept. 2009, pp. 1-12.

[43] A. Vijayaraghavan, "MTConnect for Realtime Monitoring and Analysis of Manufacturing Enterprises," Dec. 2009; http://www.systeminsights.com/, 25.12.2010.

[44] EsperTech, http://www.espertech.com/, 25.12.2010.

[45] Drools Fusion, http://www.jboss.org/drools/, 25.12.2010.

[46] M. Cammert, C. Heinz, J. Krämer, T. Riemenschneider, M. Schwarzkopf, B. Seeger, and A. Zeiss, "Stream Processing in Production-to-Business Software," Proc. of the IEEE Int. Conf. on Data Eng., pp. 168-169, 2006.

[47] A. Scheer, Business Process Engineering. Reference Model for Industrial Enterprise, 2nd Edition, Springer, 1994.

[48] M. Hammer and J. Champy, Reengineering the Corporation: A Manifesto for Business Revolution, Harper Business, 1994.

[49] N. Gronau and E. Weber, "Management of Knowledge Intensive Business Processes," In: J. Desel, B. Pernici, and M. Weske, Eds., BPM 2004, Springer, 2004, pp. 163-178.

[50] DIN 8582, Manufacturing Processes Forming - Classification, Subdivision, Terms and Definitions, Alphabetical Index, 2003.

[51] K. Kendall and J. Kendall, Systems Analysis and Design, 6th Edition, Prentice Hall, 2004.

[52] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, Fall 1996, pp. 37-54.

[53] K. McGarry, "A Survey of Interestingness Measures for Knowledge Discovery," The Knowledge Engineering Review, vol. 20, no. 1, 2005, pp. 39-61, doi: 10.1017/S0269888905000408.

[54] D. T. Pham and A. A. Afify, "Machine–Learning Techniques and their Applications in Manufacturing," Proc. IMechE Part B: J. of Eng. Manufacture, vol. 219, no. 5, 2005, pp. 395-412, doi: 10.1243/095440505X32274.

[55] M. Shahbaz, Srinivas, J. A. Harding, and M. Turner, "Product Design and Manufacturing Process Improvement using Association Rules," Proc. IMechE Part B: J. of Eng. Manufacture, vol. 220, no. 2, 2006, pp. 243-254, doi: 10.1243/095440506X78183.

[56] A. A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery," Advances in Evolutionary Computing: Theory and Applications, Springer-Verlag, Newyork, 2003, pp. 819-845.

[57] J. Han and M. Kamber, Data Mining - Concepts and Techniques, Morgan Kaufmann, London, 2001.

[58] D. Luckham and R. Schulte, Eds., Event Processing Glossary - Version 1.1, Event Processing Technical Society, 2008.

[59] D. Luckham, The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley, Munich, 2007.

[60] K. Walzer, A. Schill, and A. Löser, "Temporal Constraints for Rule-Based Event Processing", Proc. ACM first Ph.D. workshop in CIKM, Nov. 2007, pp. 93-100.

[61] IBM SPSS Modeler Professional, http://www.spss.com/, 25.12.2010.

Saliency Detection Making Use of Human Visual Perception Modelling

Cristina Oprea, Constantin Paleologu, Ionut Pirnog, and Mihnea Udrea
Dept. of Telecommunications
Politehnica University of Bucharest
Bucharest, Romania
cristina@comm.pub.ro, pale@comm.pub.ro, ionut@comm.pub.ro, mihnea@comm.pub.ro

***Abstract*—This paper proposes an algorithm for accurate detection of salient areas from a given scene. We used a complex model for the human visual system, in order to simulate the visual perception mechanisms. Human visual system modelling requires accurate knowledge about the entire visual pathways. This work focuses on the following features of the human vision: the color perception mechanism, the perceptual decomposition of visual information in multiple processing channels, contrast sensitivity, pattern masking, and detection/pooling mechanism present in the primary visual cortex. Pattern masking is considered within a complex approach, combining data from distinct dimensions. The results are shown to correlate well with the subjective results obtained from an eye-tracking experiment.**

Keywords – human visual system; saliency map; visual perception; masking; perceptual decomposition.

I. INTRODUCTION

This paper is focused on a region identification question and the regions that we are looking for are the ones having the best saliency from the perceptual point of view, as presented in [1]. The main idea is to be able to decide wich are the most important areas in a given scene, image or video frame. Such an algorithm has several applications, some of the most important beeing in the video preprocessing stage (coding), in order to optimize the compression scheme and in watermarking schemes that should hide information more effectively in images. In such applications, the characteristics and especially the limitations of the human visual system can be exploited to obtain the best performance with respect to visual quality of the output. Physiologists and psychologists have performed psycho-visual experiments aiming to understand how the human visual system (HVS) works. Engineers apply the results of those psychovisual experiments in their applications, but to do so, they use the simplified human vision models. This paper presents an attempt to integrate a such computational model of the human visual system into a tool for perceptual important areas detection. However, the experimental conditions used in the psycho-visual experiments are not representative for all types of image processing applications. Using the simplified human vision models with little knowledge regarding the applicability of these models under the new conditions limits the precision of the results.

The computational model of the human visual system that we have used is a model derived from the one introduced by [2]. This model is based on the multi-channel architecture, as first proposed by Watson in [3] who assumed that each band of spatial frequencies is dealt with by a separate channel. The contrast sensitivity function (CSF) is the envelope of the sentivities of those channels. The detection process occurs independently in any channel when the signal in that band reaches a threshold. In addition, several models proposed later, including [4] and the present work, take into consideration temporal channels as well as chromatic sensivities and orientation selectivity. The perceptual decomposition in multiple channels is then performed in both domains, spatial and temporal. The temporal channels will deal with the dynamic stimuli from the visual scene.

The paper is structured in four sections. Section II introduces the latest achievements in the area of perceptual region detection and human visual system modelling. Section III contains a detailed presentation of the proposed method, while in the final section of this paper we show that the results obtained with this algorithm are a good approximation for the perceptual regions detected with subjective experimental testing.

II. PREVIOUS WORK

Previous work related to this approach was mainly conducted in the field of visual attention modelling. Although visual assessment task in humans seems simple, it actually involves a collection of very complex mechanisms that are not completely understood. The visual attention process can be reduced at two physiological mechanisms that combined together result in the usual selection of perceptual significant areas from a natural or artificial scene. Those mechanisms are bottom-up attentional selection and top-down attentional selection. The first mechanism is an automated selection performed very fast, being driven by the visual stimulus itself. The second one is started in the higher cognitive areas of the brain and it is driven by the individual preferences and interests. A complete simulation of both mechanisms can result in a tremendously complex and time-consuming algorithm.

The process of finding the focus of attention in a scene is usually done by building feature maps for that scene, following the feature integration theory developed by Treisman [5]. This theory states that distinct features in a scene are automatically registered by the visual system and

coded in parallel channels, before the items in the image are actually identified by the observer. Independent features like orientation, color, spatial frequency, brightness, and motion direction are put together in order to construct a single object being in the focus of attention. Pixel-based, spatial frequency and region-based models of visual attention are different methods of building feature maps and extracting saliency.

The pixel-based category is represented by Laurent Itti's work concerning the emulation of bottom-up and top-down attentional mechanisms [6]. Another possibility of building feature maps is by applying different filtering operations in the frequency domain. Most common type of such filtering is done using Gabor filters and Difference of Gaussians filters. The work in [7] applies the opponent color theory and uses contrast sensitivity functions for high contrast detection. Last category of visual attention models are the region-based algorithms. In this case it is usually performed a clustering operation like region segmentation on the original image and then feature maps are computed using these clusters [8].

Regarding the HVS modelling, there have been studied and evaluated by the Video Quality Experts Group (VQEG) several video quality metrics that are using such models for the visual system. Based on a benchmark by the VQEG in the course of the Multimedia Test Phase 2007-2008, some metrics were recently standardized as ITU-T Rec. J.246 [9] and J.247 [10]. The first recommandation, J.246 presents several methods for perceptual visual quality assessment for cable television. Such networks have the advantage of permitting the transmission of some information about the reference or even a reduced bandwidth reference.

The second recommandation J.247 states a new set of methods dedicated to perceptual video quality measurement when the entire reference is available. One of those methods, PEVQ or Perceptual Evaluation of Video Quality performs a pre-processing step that extracts a region of interest from the reference and the distorted signals. All the following calculations are then performed only on that region of interest. This step is based on the observation that distortions nearest to the border are not really noticed by viewers and often get ignored. This idea can be developped into a more precise analysis and one can identify a region of interest that best fits the perceptual saliency. All further calculations can be performed for that specific area found to be closest to the human focus of attention, consuming less time and resources in the application under consideration.

III. PROPOSED ALGORITHM

Human visual system modelling requires accurate knowledge about the entire visual pathways. In present, only certain aspects of vision are well understood and so, human visual system models have been developed in order to simplify the bahaviours of what is a very complex system. As the knowledge about the real visual system improves, the model can be upgraded. Such models are used by experts and researchers in image processing, video processing and computer vision, dealing with applications related to biological and pshychological processes.

Many HVS features have their origins in evolution, since people needed to hunt for food and defend from other predators. For exemple, motion sensitivity is higher in peripheral vision with the purpose of early detection of any danger coming from wild animals. Also, motion sensivity is stronger then texture sensivity since it was crucial to scan the landscape and detect any camouflaged animals.

The model used in this work focuses on the following features of the human vision: the color perception mechanism, the perceptual decomposition of visual information in multiple processing channels, contrast sensitivity, pattern masking, and detection/pooling mechanism. In the following presentation, each feature is integrated into an algorithm processing step. The main target is to obtain at the end of the algorithm a map indicating the most salient areas from a scene. Perceptual saliency detection stands for the identification of objects, persons, visual stimuli in general that have the quality of standing out relative to neighboring items or simply beeing eye-catching. This task is similar to finding the focus of attention, that means recreating the mind's perceptual function to direct its inner awareness upon a specific target.

A. Color processing

The chromatic information from the visual scene is processed in the retinal stage according to the trichromatic theory. In the following stages of the visual pathway, specifically in the lateral geniculate nucleus, the color data is encoded according to the opponent colors theory, a technique that removes redundancy from the data stream.

At the first stage of color perception in the retina, photoreceptor cells convert the light energy into neural signals. The basic process performed by photoreceptors is absorption of photons from the field of view and signalling this information through a change in the membrane potential. This mechanism provides the subsequent cortical areas with the necessary information about the scene comprised in the field of view. There are two types of photoreceptor cells: rods and cones, and they have different functions. Rods are found primarily in the periphery of the retina and are used to see at low levels of light. Rods are not sensitive to color, only to light/dark or to black/white. Rods can function in less intense light than can the other type of photoreceptors, cone cells, and they are concentrated at the outer edges of the retina beeing used in peripheral vision. Cones are located especially in the center of the retina. There are three types of cones that differ in the wavelengths of light they absorb; they are usually called short or blue (S), middle or green (M), and long or red (L). Cones are used to distinguish color at normal levels of light.

In later stages of visual information processing, the color is to be coded differently. From the three primaries given by cones and the intensity given by rods, the color is eventually encoded as one luminance channel (magnocellular cells from the lateral geniculate nucleus - LGN) and two crominance

channels: one for red-green cones (parvocellular cells in LGN) and another one for blue-yellow connes (koniocellular cells in LGN).

The color processing block in our algorithm is conducting a conversion from the usual YCbCr color-space to an opponent color space, similar to the one discovered at the LGN level. The resulting color components are: W-B for white-black, R-G for red-green, and B-Y for blue-yellow. These opponent colors can be associated to a luminance signal and two color difference signals. The colors selected are not random, they are considered opponent because under normal circumstances there is no hue one could describe as a mixture of opponent hues [11].

In order to obtain those components, the trichromatic values (RGB computed from YCbCr) undergo a power-law nonlinearity to counter the gamma correction used to compensate for the behaviour of conventional CRT displays. In LCD displays, the relation between the signal voltage and the intensity is very nonlinear and cannot be described with gamma value. However, such displays apply a correction onto the signal voltage in order to approximately get a standard γ=2.5 behavior.

The linear RGB values produced are then converted to responses of the L, M and S cones on the human retina, based on the spectral absorbtion measured for these cells. This conversion is performed in two steps: first, RGB color space is converted to CIE XYZ color space; second, from XYZ components will be computed the LMS values. For the first transformation we have used a matrix defined in ITU-R Rec. BT.709-5 [15]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.358 & 0.180 \\ 0.213 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

The responses of the L, M, and S cones from the human retina are computed according to CIECAM02, the most recent color appearance model ratified by CIE Technical Committee (International Commission on Illumination) [16]:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.7328 & 0.4296 & -0.1624 \\ -0.7036 & 1.6975 & 0.0061 \\ 0.0030 & 0.0136 & 0.9834 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (2)$$

There is no unanimity of opinion regarding the particular values of the coefficients used in those transformations and several papers still use the classical von Kries transformation. We prefered the transformation matrix proposed in the latest standard published by ITU-R since it comes from more recent studies and measurements.

Knowing the L, M, S cones absorbtions rates, the convertion to an opponent color space becomes possible due to the transformation matrix proposed by Poirson and Wandell [16]. The same transformation matrix has also been used by Winkler in his Perceptual Distortions Metric [2]:

$$\begin{bmatrix} W-B \\ R-G \\ B-Y \end{bmatrix} = \begin{bmatrix} 0.990 & -0.106 & -0.094 \\ -0.669 & 0.742 & -0.027 \\ -0.212 & -0.354 & 0.911 \end{bmatrix} \cdot \begin{bmatrix} L \\ M \\ S \end{bmatrix} \quad (3)$$

The color space proposed by Poirson and Wandell was developed aiming to completely separate the color processing from the pattern perceptual processing. Keeping apart the color from the pattern makes easier to simulate the mechanisms in the human vision.

The opponent color space agrees with the color processing at higher levels in the human brain, especially in the cortical area called V1. This type of color encoding de-correlates the signals coming from the retina and removes redundancy. In fact, in area V1 it has been proven to exist two types of double-opponent cells: red-green and blue-yellow. Red-green cells confront the relative amounts of red-green in one part of a scene, with the amount of red-green in a neighboring part of the scene; such cells respond best to local color contrast (red next to green).

B. Multi-channel decomposition

The multi-channel decomposition is performed according to a theory that explains the visual perception of a scene including multiple visual stimuli: each feature from the input scene is processed separately. Many cells in the human visual system and mainly in the visual cortex have been proven to be selectively sensitive to certain types of signals such as patterns of a particular frequency or orientation.

The visual cortex is made from the combination of several areas: V1 (or primary visual cortex), V2, V3, V4, and V5. Neurons in the visual cortex respond to visual stimuli that appear within their receptive field by sending action potentials. The receptive field of one neuron is the region within the entire visual field which causes a response from that neuron. Each neuron responds best only to a subset of stimuli within its receptive field. This mechanism is neuronal tuning. First visual areas (for exemple V1 area) have neurons with simpler tuning that will respond to stimuli falling in their receptive fields such as vertical lines or textures with particular spatial frequencies. In later visual areas, neuronal cells have complex tuning that is much more complicated cu simulate. For instance, a neuron in the inferior temporal cortex may only react when a certain face appears in its receptive field.

During the experiments regarding the primary visual cortex, it has been noticed that the tuning properties of V1 neurons differ greatly over time. Evidence shows that there are at least two temporal mechanisms that affect neuronal responses in V1. The overall functioning of V1 can be thought of tiled sets of selective spatiotemporal filters. This is why the multi-channel decomposition splits the input into a number of channels, based on the spatio-temporal mechanisms present in area V1 from the visual cortex. In theory, these filters together can carry out neuronal processing of spatial frequency, orientation, motion, direction, speed (thus temporal frequency), and other spatiotemporal features.

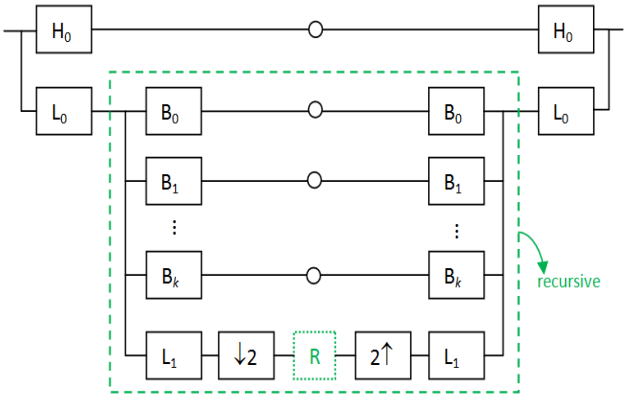


Figure 1. Simoncelli's steerable pyramid. Downsampling by a factor of 2 and upsampling by 2 are used. The recursive construction of the pyramid is achieved by inserting a copy of the diagram contents enclosed by the dashed rectangle at the location of the block "R".

Temporal mechanisms are modeled with a perceptual decomposition in the temporal domain. We used two filters for two temporal mechanisms: the sustained and transient mechanisms, the same filters used in [2] and proposed by Fredericksen and Hess [17]. Finite impulse response (FIR) filters with linear phase are computed by means of a least-squares fit to the normalized frequency magnitude response of the corresponding mechanism as given by the Fourier transforms of $h(t)$ and $h''(t)$, the second derivative of $h(t)$, from the following equation:

$$h(t) = exp[-(5\ln(t/\theta))^2] \quad (4)$$

The sustained mechanism is implemented by a low-pass filter, while the transient mechanism – by a band-pass filter. Both FIR filters are applied only to the luminance channel in order to reduce computing time. This simplification is based on the fact that our sensitivity to color contrast is reduced for high frequencies.

Spatial mechanisms are modeled by means of a steerable pyramid decomposition [12], first proposed by Simoncelli. In this linear decomposition, an image is subdivided into a collection of subbands localized in both scale and orientation. Similar multiscale transforms have often been used in image processing and image representation. For exemple, the wavelet transform was proven usefull in applications were scalable video coding was needed.

The scale tuning of the filters is constrained by a recursive system diagram, as illustrated in Fig. 1. The left part of the diagram is called the analysis filter bank, while at the right side, the synthesis filter bank performs the reconstruction of the original image. The orientation tuning is constrained by the property of steerability, which means that the transform is shiftable in orientation. A set of filters form a steerable basis if :

(i) they are rotated copies of each other and

(ii) a copy of the filter at any orientation may be computed as a linear combination of the basis filters.

The pyramid's algorithm itself is based on recursive application of two types of operations: filtering and subsampling. First, the input signal or the original image/frame is divided into a low-pass and a high-pass portions. The low part will be further subdivided into bandpass portions and another low-pass one; each of the bandpass filters select features having distinct orientations. The last low-pass portion obtained is subsampled by a factor of 2 and the algorithm will be repeated in recursive cascades. The bandpass divisions are not subsampled in order to avoid aliasing, while for the subsampled low-pass subimage, the aliasing issue is prevented by using low-pass radial filters especially designed.

In addition to having steerable orientation subbands, this transform can be designed to produce any number of orientation bands, k. The resulting transform will be overcomplete by a factor of $4k/3$, meaning that the coefficient output rate is greater then the input signal sample rate. Note that the steerable pyramid retains some of the advantages of orthonormal wavelet transforms, but improves on some of their disadvantages (e.g., aliasing is eliminated; steerable orientation decomposition). One obvious disadvantage is in computational efficiency: the steerable pyramid is substantially overcomplete.

Six sub-band levels with four orientation bands each plus one low-pass band are computed; the bands at each level are tuned to orientations of 0, 45, 90, and 135 degrees, as illustrated in Fig. 2. The same decomposition is used for the W-B, R-G and B-Y channels, meaning that all color channels go through the same steerable pyramid transform. This approach agrees with the primary visual cortex architecture regarding color, spatial frequency, and orientation processing.

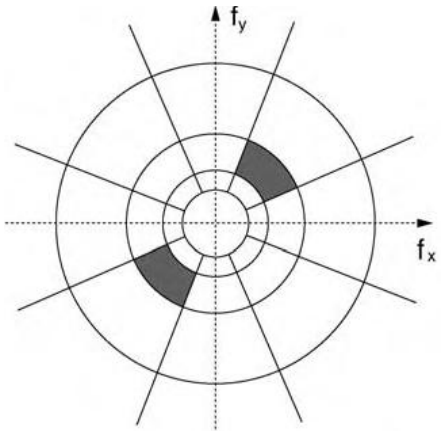
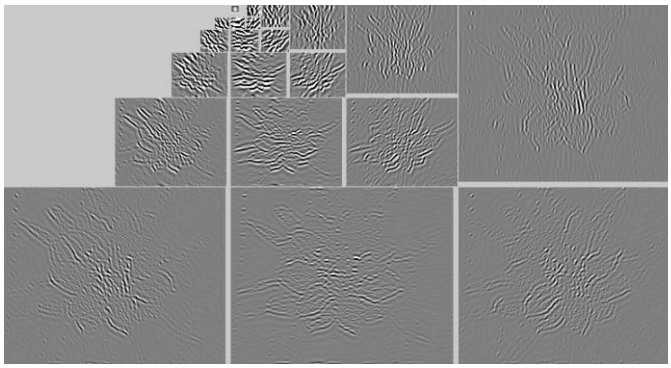


Figure 2. Spatial frequency plane partitioning in the steerable pyramid transform. The gray region indicates the spectral support of a single sub-band oriented at 45 degrees [2].

At this point of the algorithm, the input image is subjected to the steerable pyramid transform and the result is illustrated in Fig. 3 for a set of two images. The first image illustrates a flower who's petals have radial disposition, thus containing all orientations. In the output of

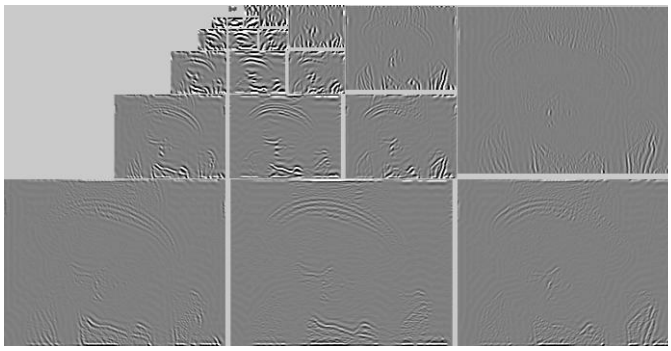
the steerable pyramid transform, at the first decomposition level it is easy to recognize the extracted feature's orientations.

a)



b)

c)



d)

Figure 3. a) and c) are original test images; b) and d) are the outputs from the steerable pyramid transform (four orientations and five levels decomposition).

C. Contrast sensitivity

Contrast is a visual property that makes an object distinguishable from neighboring elements or background. The human visual system is more sensitive to contrast than to absolute luminance and the human eye itself is designed to react only to luminance variations.

Researchers built contrast sensitivity functions from experimental measurements and beside the classic luminance contrast given by white/dark association, we now have color contrast sensitivity curves for the two chromatic channels: red-green and blue-yelow [18], [19] as it can be seen in Figure 4. The contrast sensitivity function shows a typical band-pass shape peaking at around 4 cycles per degree with sensitivity dropping off either side of the peak, meaning that human vision is most sensitive in detecting contrast differences occurring at 4 cycles per degree. The high-frequency cut-off represents the optical limitations of the visual system's ability to resolve detail and is typically about 60 cycles per degree. Tipically, our sensitivity to color contrast is reduced for high frequencies.

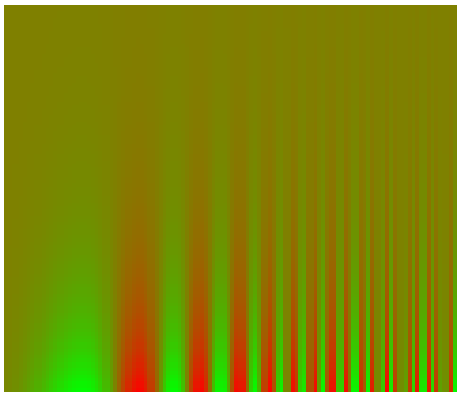


Figure 4. The contrast sensitivity function for the red-green channel is the envelope of the visible gratings.

Next step after the temporal and spatial decomposition is a shortcut in computation efficiency. Instead of pre-filtering the W-B, R-G and B-Y channels with their respective contrast sensitivity functions, which is the accurate approach, we searched for a set of weighting factors for each channel. The weights were determined intending to obtain a filter set that approximates the spatio-temporal contrast sensitivity of the human visual system. It was preferable since it conducts to a more simple implementation and the simulation time improves.

D. Masking

Visual masking is a perceptual phenomena that stands for the reduction or the elimination of the visibility of one brief stimulus called the "target" due to the presentation of a second brief stimulus, called the "mask". Within the framework of quality assessment it is helpful to think of the distortion or the coding noise as being masked by the original image or sequence acting as background. Masking

explains why similar coding artifacts are disturbing in certain regions of an image while they are hardly noticeable in others. In order to be possible for visual masking to appear, both the target and the mask must be briefly presented, less then 50ms.

Our human visual system model implements both intra-channel and inter-channel masking. Masking is known to be stronger between visual stimuli of the same type (located in the same decomposition channel), so called intra-channel masking. This type of visual masking appears for a pair target-mask that have the same characteristics: belong to the same frequency band, same orientation, and even identical chromaticity. But masking also happens, at a lesser extent, between stimuli coming from different channels, beeing called inter-channel masking. We approached the masking perceptual process as a question of multiple excitations and inhibitions flows in the cortical pathways. For a neuron's excitation stronger than the associated inhibition from other neurons, we obtain the evidentiation. The opposite phenomenon, an excitation weaker than corresponding inhibitions will emulate the perceptual masking. An accurate modelling of evidentiation and masking operations will bring foreward salient features and objects from the input image.

In neural networks, the neuron's excitation or inhibition can be simulated with the following linear equation:

$$y_j^N = \sum_i g\left(w_{ji} x_i + c_j\right) \tag{5}$$

where the output or the response of the *j*-th neuron is given by all inputs to that neuron, indexed by *i* and denoted as *x*, weighted by the coefficients *w* according to that neuron's specialization. Excitation appears for a positive weight, while inhibition follows a negative weight.

Our model takes into consideration the excitatory behaviour of specialised neurons inhibited by a pool of responses from other nervous cells in the visual cortex. Instead of a linear model, we adopted a nonlinear model where the weights were removed, thus eliminating the problem of choosing their values. The excitation is modeled by a power-law nonlinearity, where the input *x* is raised at power *p*. The inhibition follows the same modelling rule having an exponent *q*.

$$y = \frac{x^p}{c + h_1 * x^q + h_2 * x^q} \tag{6}$$

Equation 6 illustrates that the excitatory behaviour can be modelled by means of a power-law nonlinearity with exponent *p* greater then the inhibitory exponent *q*. The numerator models the excitation and *x* is a coefficient from the perceptual multi-channel decomposition. Such a coefficient comes as output from one of the filters in the filter bank that comprises the steerable pyramid. Therefore, *x* is a coefficient that carries information about a feature in the input image having precise characteristics: spatial frequency, color, and orientation. The denominator contains a constant *c* that prevents division by zero and two convolutions: h_1 represents a gaussian pooling kernel for coefficients from the same decomposition channel, while h_2 is another gaussian pooling kernel for different channels interactions. This approach has proven to be more accurate then using a single pooling kernel for all coefficients. In the inhibitory path, filter responses are pooled over different channels by means of two convolutions, combining coefficients from the dimensions of space and orientation.

E. Detection

The information coded in multiple channels within the VI area of the visual cortex is integrated in the subsequent cortical areas. This process can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling.

Then, the steerable pyramid is reconstructed only for the luminance channel.

a) b) c)

Figure 4. a) Original image "Lena"; b) Saliency map obtained with our algorithm; c) Importance map obtained with TBQM metric [13]. The brighter pixels have higher saliency/ perceptual importance.

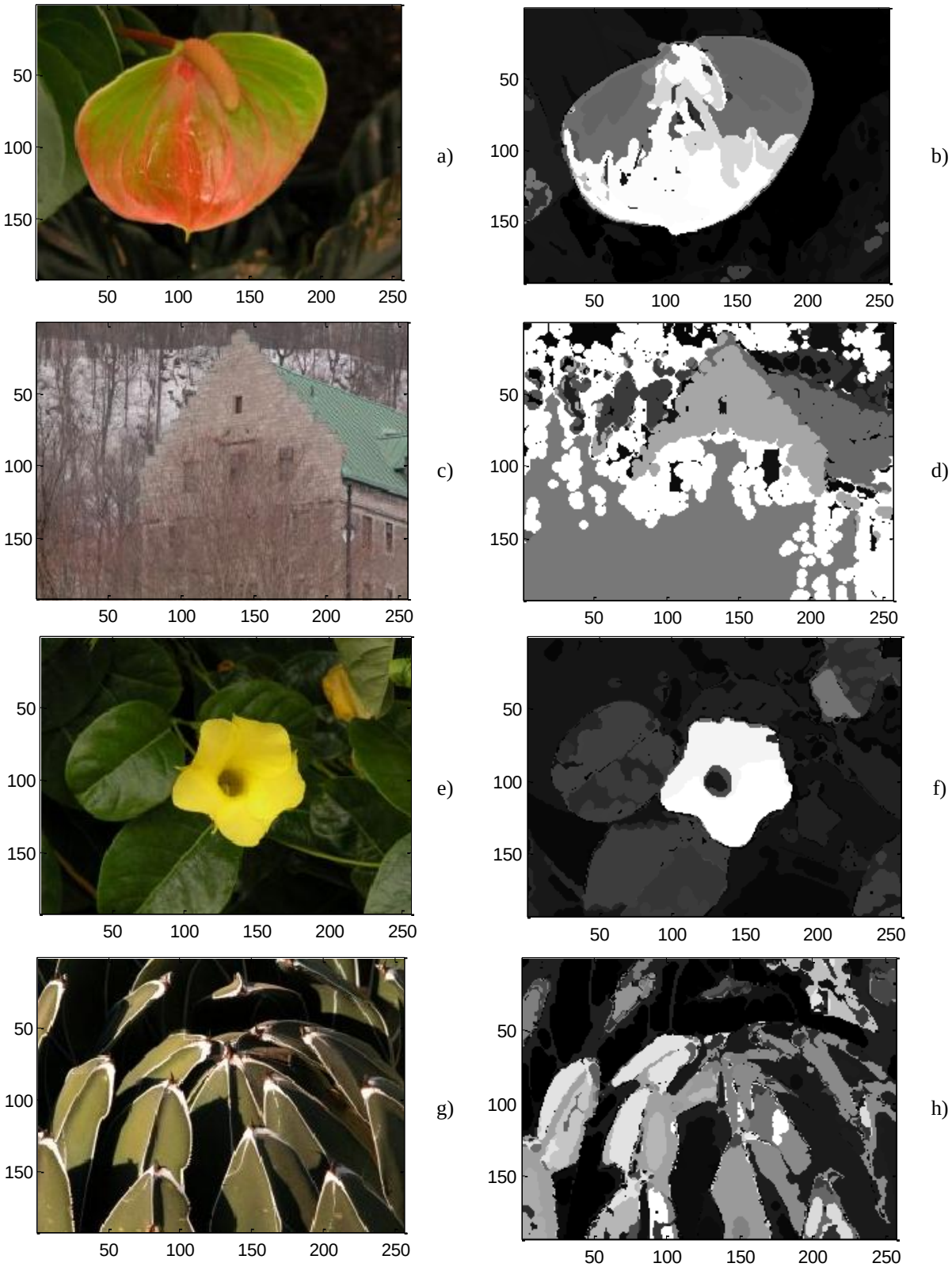


Figure 5. a), c), e) and g) are original test images from the eye-tracking experiment database [14]; b), d), f) and h) are saliency maps obtained with the algorithm described previously. The brighter areas have a stronger perceptual importance, while the dark zones designate features without saliency.

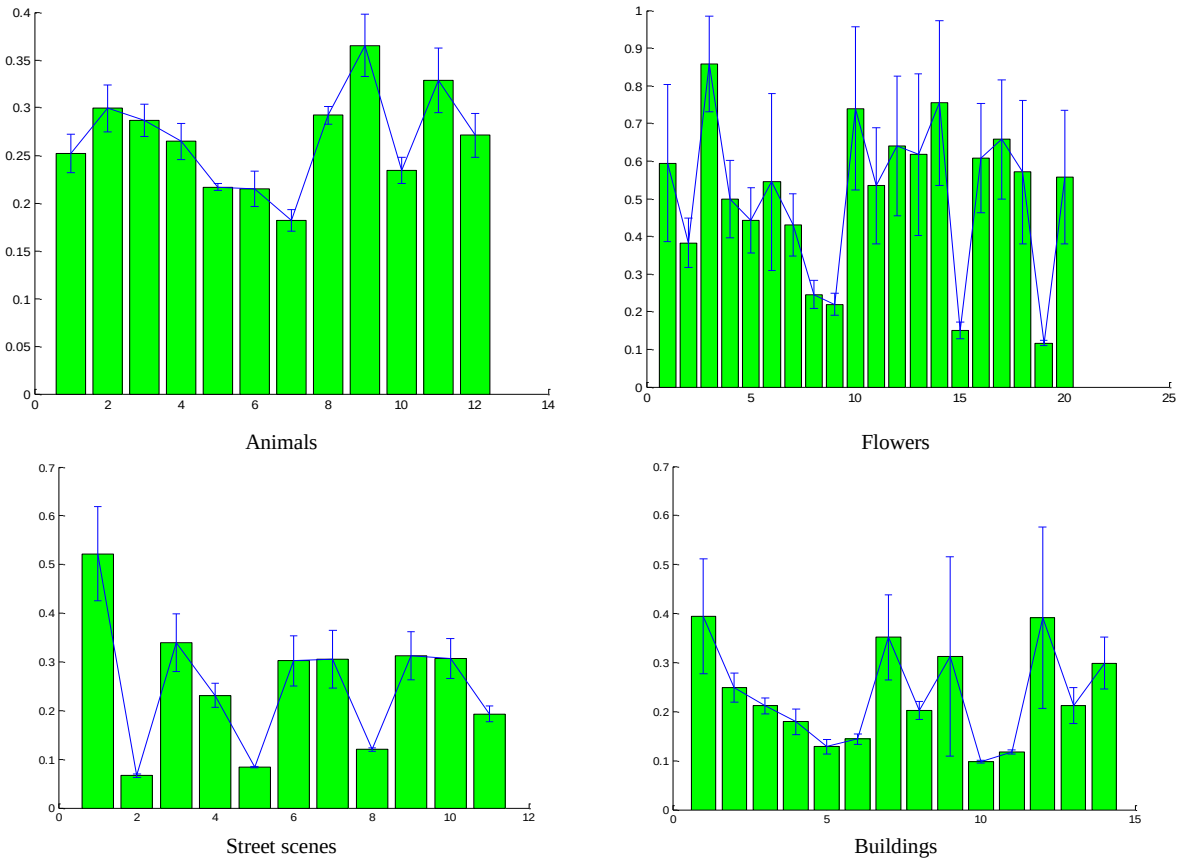


Figure 6. Correlations between human fixation maps from the eye-tracking experiment and objective saliency maps.

Our model of the human visual system uses at this stage a thresholding operation for the set of coefficients *y* resulted in the previous processing stage. Threshold values are model constants determined by experimental simulations, taking into consideration the saliency related to multiscale representation. The chromatic information was already included by the inter-channel masking processing. The resulting image pixel values are brought to a subunitary domain and they represent the saliency contained in the original image.

IV. RESULTS

For comparison of our results with one of the existing models, we provided the output of saliency maps by [13] in Figure 4. The model proposed in this paper is inclined towards totally eliminating the noninteresting areas because of being strict in selection. This tendency confines the exploration to a limited number of spots and the probability to skip a moderately prominent object in a visual search turns high. The method for importance map construction by [13] shown to be unable to discriminate the saliency of naturally prominent colors and also it does not consider the global context in the given scene. The proposed method has eliminated such weaknesses by incorporating theory of colors into the model and by including the influence of local and global neighborhood on the saliency of objects. In Figure 5 there are presented four test images and their objective saliency maps determined with our algorithm.

The saliency maps are also correlated with the subjective results obtained for a 29 test images database, containing eye-tracking data [14]. Such data are highly accurate due to the experimental setting and the testing subjects carrefully selected. Eye-tracking data result in the only subjective saliency maps that can be used for comparison with objective methods. In order to compare the saliency maps with the human data, we used a correlation method proposed in [14]. The value of comparison is given by the correlation coefficient ρ:

$$\rho = \frac{\sum_{x,y}(OM(x,y) - m_{OM})(SM(x,y) - m_{SM})}{\sqrt{\sigma_{OM}^2 \sigma_{SM}^2}} \quad (7)$$

where $OM(x,y)$ is the objective map, $SM(x,y)$ is the subjective map, and m, σ^2 are the mean and the variance of the values from these maps. A positive correlation coefficient indicates similar structure in both maps. Our objective maps result in correlation coefficients greater then those obtained with TBQM metric in 73% cases. In Figure 6 are illustrated the correlation coefficients for four

different types of images: natural scenes with animals, natural scenes with flowers, street scenes with peoples and cars and finally, building scenes. All the images come from the database provided by [14]. Each bar represent the mean correlation coefficient for the computed correlations between the 31 human fixation maps and our saliency map. The error bars give the confidence intervals.

V. CONCLUSIONS

The model proposed exploits spatiotemporal information and provides an efficient preprocessing step (salient spatiotemporal event detection) that will limit the application of high-level processing tasks to the most salient parts of the input. Our model simulates only the behaviour of the primary visual cortex (V1), which is necessary for conscious vision. As future work, the algorithm will be upgraded with emulations of the superior extrastriate visual cortex areas that will replace the final detection operation performed in the current work.

ACKNOWLEDGMENT

This work was supported by the UEFISCDI grant PN-II-RU-TE no. 7/5.08.2010.

REFERENCES

[1] C. Oprea, C. Paleologu, I. Pirnog, M. Udrea, "Saliency Detection Based on Human Perception of Visual Information" pp.96-99, 2010 Sixth Advanced International Conference on Telecommunications, 2010.

[2] S. Winkler, “Digital Video Quality Vision Models and Metrics”, John Wiley & Sons, 2005, ISBN 0-470-02404-6.

[3] A.B. Watson, “The cortex transform: Rapid computation of simulated neural images”, Computer Vision, Graphics, and Image Processing, 1987, 39,3, pp. 311-327.

[4] M. A. Masry and S. S. Hemami, “A metric for continuous quality evaluation of compressed video with severe distortions”, Signal Proc.: Image Communication, 2004, vol. 19, (2), pp. 133-146.

[5] A. M. Treisman and G. Gelade, “A feature-integration theory of attention”, Cogn. Psychol., vol. 12, pp. 97-136, 1980.

[6] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention”, IEEE Trans. Image Process., 2004, vol.13, (10), pp. 1304 – 1318.

[7] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, “A coherent computational approach to model bottom-up visual attention”, IEEE Trans.Pattern Anal. Mach. Intell., 2006, vol. 28, pp. 802-817.

[8] C. Oprea, I. Pirnog, C. Paleologu, M. Udrea, “Perceptual Video Quality Assessment Based on Salient Region Detection”, Proceedings of AICT 2009, May 2009, pp. 232-236.

[9] ITU-T Rec. J.246, “Perceptual audiovisual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference”, 2008.

[10] ITU-T Rec. J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”, 2008.

[11] M. Nida-Rumelin, J. Suarez, "Reddish Green: A Challenge for Modal Claims About Phenomenal Structure". Philosophy and Phenomenological Research 78: 346. doi:10.1111/j.1933-1592.2009.00247.x, 2009.

[12] Q. Wu, M. A. Schulze, K. R. Castleman, “Steerable Pyramid Filters for Selective Image Enhancement Applications”, Proceedings of ISCAS ‘98, 1998.

[13] G. Zhai, W. Zhang, X. Yang, Y. Xu,” Image quality metric with an integrated bottom-up and top-down HVS approach”, IEE Proc.-Vis. Image Signal Process., Vol. 153, No. 4, August 2006.

[14] G. Kootstra, A. Nederveen, B. De Boer, “Paying attention to symmetry”, Proc. British Machine Vision Conference, UK, 2008.

[15] ITU-R Rec. BT.709-5, “Parameter values for the HDTV standards for production and international programme exchange”, ITU, Geneva, Swiss, 2002.

[16] A. B. Poirson, B. A. Wandell, “Pattern-color separable pathways predict sensitivity to simple colored patterns”, Vision Research vol. 36 (4), pp. 515-526, 1996.

[17] R. E. Fredericksen, R. F. Hess, “Estimating multiple temporal mechanisms in human vision”, Vision Research, vol. 38(7), pp.: 1023-1040, 1998.

[18] F. W. Campbell, J. G. Robson, “Application of Fourier analysis to the visibility of gratings”, Journal of Physiology, 1968.

[19] E. Peli, “Contrast in complex images”, Journal of the Optical Society of America, vol. 7 (10), pp.2032-2040, 1990.

Core-Body Temperature Acquisition Tools for Long-term Monitoring and Analysis

João M. L. P. Caldeira[1,2], Joel J. P. C. Rodrigues[1], José A. F. Moutinho[3], Marc Gilg[4], and Pascal Lorenz[4]
[1] *Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal*
[2] *EST, Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal*
[3] *Health Sciences Faculty, University of Beira Interior, Covilhã, Portugal*
[4] *IUT, University of Haute Alsace, Colmar, France*
jcaldeira@it.ubi.pt, joeljr@ieee.org, jafmoutinho@fcsaude.ubi.pt, marc.gilg@uha.fr, lorenz@ieee.org

***Abstract*— The detection of fertile and ovulation periods may be performed by women's body temperature variations. These variations are more accurate if a core-body temperature for their detection is used. Previous medical studies concluded that the use of skin temperature could be influenced by environmental conditions. Since the increasing of the body temperature in this period is only about 0.5 ºC, it is crucial that measurements should be the most accurate as possible. Due to the lack of solutions to realize that in order to measure and analyze the core-body temperature, this paper presents a system to capture, display, and monitoring core-body temperature. It is considered a hardware solution (sensor) to be placed inside cervix and a computer application to communicate and gather the collected data by the sensor. Bluetooth is used to perform the communication between a computer and the sensor. The system evaluation is performed by a medical team in several volunteer women. Furthermore, the collected data by the sensor may be used to study the relation between temperature variations and women health conditions.**

***Keywords*— Biosensor; Wireless sensor network; Biofeedback; e-Health; Temperature Monitoring.**

I. INTRODUCTION

New technologies applied to healthcare and biofeedback improves the traditional way of medical procedures. Recent publications [1-3] report continuous evolution and progress of new biosensors for healthcare and biofeedback. These sensors became indispensable in the daily routine of medical staff where they have the capability for helping medical procedures and healthcare. Nowadays, biosensor systems are powerful available instruments in diagnosis, controlling, monitoring and prevention of some diseases [4, 5]. In some cases, they also became an essential instrument for heal support [6-8].

The evolution of these biosensors offers a new range of the infinity possibilities for applications they can provide. The miniaturized size of these nodes turns these systems more easy to use, in a comfortable way. They can access to inside-human body places that were difficult to reach and non comfortable for patients, using traditional methods [9]. Advantages of these systems and the great interest of medical community turn this research area as an important topic.

The human body temperature is one of the most controlled bio-parameters because it reflects some health conditions through its variations. Monitoring this human parameter may improve healthcare on patients suffering from pathologies that could be controlled by body temperature regulations. In women this parameter is also correlated with fertility stages. The increasing of regular core-body temperature by about 0.5 centigrade degrees (ºC) probably indicates the occurrence of a fertility period. Therefore, monitoring this parameter becomes an excellent method to predict this period [13]. The acquisition of core-body temperature in women is crucial for the validity of the monitoring procedure. Digital thermometers are highly used for temperature measurements acquisition. This method is very inappropriate for active women that have to measure their core-body temperature at specific hours, in order to establish standard patterns. Therefore, this method also could lead to wrong measurements caused by rapid execution of the procedure and inappropriate handling of thermometers. The use of standalone systems could suppress the women intervention to collect this parameter. Although, these systems improve the quality of the collected values because they are less prone to bad handling that may lead to wrong measures.

This paper proposes an integrated system for long-term data acquisition, processing and analysis of cervix women's temperature. The system comprises three modules. First module is the temperature sensor (thermistor) it self. It is placed inside women body, close to the cervix. The second module is the processor unit responsible for data acquisition and long-term collection of the temperature values. Finally, the third module is a computer application software, used to operate the biosensor, and for representation and control of the intra-body temperature measured values.

This study is a joint work with physicians from the Health Sciences Faculty of the University of Beira Interior, Covilhã, Portugal. This new biosensor allows the execution of exploratory studies to increase the knowledge of female intra-vaginal physiological behavior. To perform this study, medical team wants to monitor and analyze the intra-vaginal temperature during the female menstrual cycle. Furthermore, they will use this new system for the following applications: preterm labor prevention, detection of pregnancy contractions, anticipation and monitoring of the ovulation period (for both natural contraception and *in vitro* fertilization purposes), effectiveness of some gynecology therapeutics, and supporting the discovery of new possible contraception methods. These system applications will be conducted taking into account previous medical studies where this physiological parameter (the temperature) is correlated with several human phenomena [10-12], including female fertility issues [13]. This confirms the importance of the contribution presented in this paper.

The remainder of the paper is organized as follows. Section II elaborates on some available projects carried out on this topic. Section III describes the new biosensor and presents the system architecture and the construction of the biosensor firmware including the used communication mechanisms is presented in Section IV. In section V, the intra-vaginal temperature monitoring computer application is presented, focusing in its construction and validation. The biosensor validation and results are presented in Section VI. Finally, Section VII concludes the paper and points further research directions.

II. RELATED WORKS

Research on healthcare has been striving to find relationships between core body temperature at female genitals and certain health conditions, such as ovulation period. A study presented in [10] concludes that exists a correlation between covert attention and basal temperature changing during the menstrual cycle phase based on 22 adult females and proved the importance of basal (intra-vaginal) temperature. In this study, a traditional way was used for temperature measurement. However, automatic measurements and analysis of intra-vaginal temperature readings in an unobtrusive and efficient way are desirable.

One of the earliest known projects for vaginal temperature measuring was presented in 1994 [14] and 1996 [15]. The system needs a permanent radio-frequency connection to a computer. The computer receives all the temperature measures and collects them. The sensor itself cannot store the measured values. This limitation imposes the sensor must be near to a personal computer and it difficult the mobility of the monitored woman. The use of mobile systems frees women to follow their regular daily life always under monitoring. Although, if a sensor itself cancollect the temperature measures in long-term way, it is dispensable a permanent connection to any device which improves the sensor's battery lifetime.

Another study uses a radio pill created for astronaut use, to access internal body temperature on athletes, and take measures to cool them down, avoiding excessive fatigue [16]. However, such pill-based solution introduces issues on pill elimination, and the biosensor cannot be reused again.

Other medical studies about developing integrated systems to acquire and monitor physiological parameters, including body temperature [7, 17]. These systems only measure the skin temperature. From [11] one can conclude that skin temperature cannot reflect the basal body temperature as it changes depending on the environment temperature. The AMON research team included a temperature sensor on their wearable system (AMON) [7] to study a possible correlation between the skin temperature readings by the sensor and the core body temperature. They concluded that skin temperature could be influenced by the environment conditions. Therefore, they could not show any correlation between skin temperature and core body temperature.

DuoFertility project [17] created a system to predict women fertile period. This system bases its prediction on the measurement of skin temperature. During fertile period, the variation of women core body temperature occurs around 10-14 days of menstrual cycle. It only changes about 0.5 degree Celsius [13]. Thus, trying to get core body temperature by measuring skin temperature could lead to wrong interpretations and conclusions.

In [18], the authors presented a system using UHF radio telemetry to measure the vaginal temperature and monitor the temperature. Another approach, presented in [19], the author proposed a highly accurate system where a capsule shaped sensor measures the central body temperature. This sensor can be ingested or inserted rectally such that it will transmit core body temperature continuously.

A method for detecting and predicting the ovulation and the fertility period in female mammals is described in [20]. This method provides information relating the fertility of females mammals. It comprises the following steps: (*i*) takes multiple temperature readings from the female mammal during an extended period; (*ii*) identifies and disregards temperature readings having one or more characteristics of irrelevant or faulty data; (*iii*) obtains one or several representative temperature values for the extended period; (*iv*) repeats steps from (*i*) to (*iii*) over multiple extended periods; and (*v*) analyzes the representative temperature values obtained over multiple extended periods for one or more patterns in the representative temperature values. It

indicates or predicts the ovulation in order to provide information related to the fertility of the female mammal. This method only describes a procedure to get temperature measurements for fertility purposes in female mammals and not really a hardware system that may allow this operation.

Next sections present, in detail, the construction of the new intra-vaginal temperature biosensor, and the corresponding application and communication system. This system deploy results on a personal computer for monitoring and further analysis.

III. INTRA-BODY BIOSENSOR

Sensing is fundamental to all sensor networks, and its quality depends from many factors such as size and used materials.

Body sensors measure core body temperature, ambulatory blood pressure, blood oxygen etc. As the accurate measure of core body temperature is highly preferred in numerous medical applications, intra-body biosensor is required. The main challenge is the construction of a novel intra-body biosensor for intra-vaginal temperature monitoring. Furthermore, the intra-body biosensor must consider the sensitivity of the body area, critical for the comfort of the user on a daily basis.

This proposal falls in the conception of a novel biosensor device to measure intra-vaginal temperatures and continuously gather their measurements for further analyses purpose. To access and analyze all data collected by this intra-body biosensor, a new application was also developed.

This biosensor uses a SHIMMER platform (Sensing Health with Intelligence Modularity, Mobility and Experimental Reusability). It is a wireless sensor platform designed by Intel and can be used as the central processor unit of the biosensor network. This platform has a Texas Instruments MSP430 CPU (8MHz), a class 2 Bluetooth radio communication, an IEEE 802.15.4 Chipcon wireless transceiver (2.4GHz), a 3-Axis Freescale accelerometer, a MicroSD slot for up to 2Gbytes, an integrated Li-Ion battery management and some extensions to append new features and functionalities mounted in a very small form factor (2 x 4.5 cm) not larger than a thumb size. This platform is an ideal hardware for this work due to its small size, large data storage capacity and communication features. However, there is no temperature sensor installed on SHIMMER platform. In order to get temperature readings, a temperature sensor must be integrated on SHIMMER. Thus, the MA100 thermistor from GE Industrial Sensing is used on this solution. Its sensitivity ranges from 0 to 50 degree Celsius, size is 0.762 x 9.52 mm, and is created for biomedical applications. MA100 is connected to the SHIMMER with a flexible cable.

Due to its bulky size, is not possible to place SHIMMER inside the female cervix. Therefore, only MA100 thermistor is introduced inside the vagina and sends measurements (in voltage) to SHIMMER, using a flexible cable. The SHIMMER unit stays outside the women body and could be placed anywhere if the cable is long enough (~ 80 centimeters). As the cable is very flexible and difficult to handle inside the female cervix, a tampon-like enclosure was created for the biosensor. This solution allows that women can easily introduce the tampon-like enclosure with temperature sensor inside body because it can be used as ordinary tampons.

Figure 1 presents the proposed system architecture. The temperature sensor (MA100) is placed inside cervix, while the processor unit (SHIMMER) remains outside. Bluetooth performs the communication between this device and a computer.

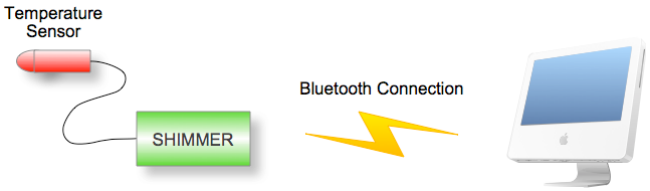


Figure 1. System architecture.

To ensure the acquisition and storage of the read temperature values, a SHIMMER firmware was developed. This software, running in continuous mode, is waiting for personal computer commands over a Bluetooth connection. Once a command is arrived, SHIMMER analyzes it and proceeds in accordance with it. The available commands in SHIMMER firmware are the following: start collecting the temperature values and save them on a microSD, stop collecting, turn on a red led (for debugging purposes), programming the interval between temperature readings and send all the recorded data in microSD to a personal computer application.

Valuable results can only be collected if a correspondence between the measured temperature and the exact time it was acquired may be identified. SHIMMER has a local time clock starting on the startup time, however it does not have a global time clock. In order to provide a global time clock on SHIMMER, when a computer performs a start collection command, it also sends its time and date to SHIMMER (assuming that computer clock is synchronized with global time clock). This information is used on SHIMMER, regarding its local time, to calculate the global time clock associated with each measure.

The proposed biosensor is presented in Figure 2. Figure 2(a) shows the MA100 thermistor in its enclosure. The SHIMMER platform and the external extension where MA100 is connected may be seen in Figure 2 (b).

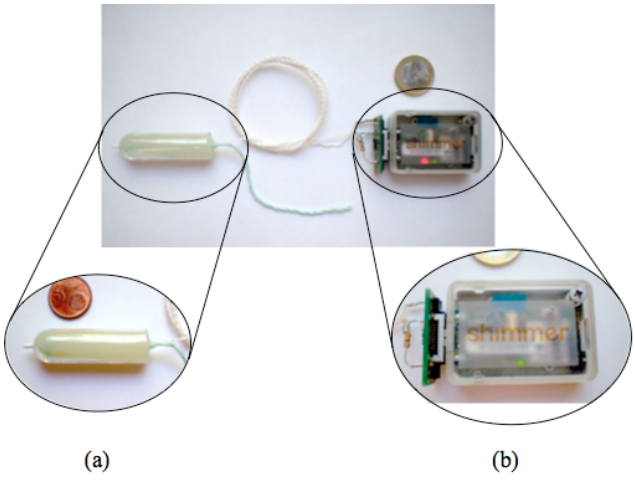


Figure 2. Biosensor for intra-vaginal temperature collection.
(a) Temperature sensor (MA100) with enclosure; (b) SHIMMER platform.

IV. COMMUNICATION TOOLS

Although SHIMMER is a powerful biosensor platform with all the above-described features, it has very limited resources in terms of computation and is very dependable from a power battery with no long lifetime. In order to ensure increase of the battery lifetime, only temperature readings from SHIMMER are gathered, instead of collecting and processing them. Temperature readings are transmitted to a computer and collected by an application for further processing and analysis.

Bluetooth performs the communication between SHIMMER and a personal computer. Wireless communications seems to be more realistic than other wired alternatives, taking into account patients comfort and operation simplicity by medical staff. Like any body area sensor network, it is unique and it attempts to restrict the communication radius to the body's periphery. Limiting transmission range reduces a node's power consumption, decreases interference, and helps privacy maintenance.

The connection between the biosensor and a computer is only available if the sensor is in Bluetooth's connection range to the computer, as expected. Because of that, an effective monitoring of the temperature values cannot be performed if the sensor is out of range. Therefore, in case of communication failure between SHIMMER and a personal computer, SHIMMER only collects temperature and saves temperature readings on its local microSD. Then, it can transmit them when a Bluetooth connection is active. This procedure prevents unnecessary use of power to perform the communication and increases battery's lifetime. Figure 3 shows a diagram of Bluetooth data transmission presenting the procedure performed in case of existence of an active Bluetooth connection, or not.

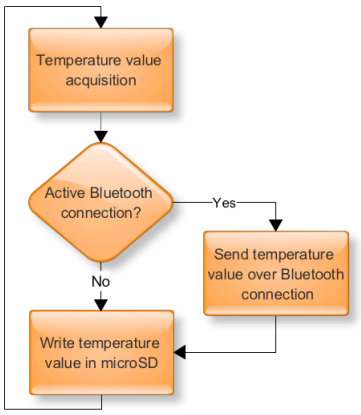


Figure 3. Diagram of Bluetooth data transmission, existing an active connection or not.

To operate with SHIMMER, several commands were implemented. The computer sends these operating commands to SHIMMER, which in turn, sends information and data to the computer. SHIMMER always waits for computer commands over a Bluetooth connection. Once SHIMMER receives a command, it proceeds accordingly. The operating commands available on SHIMMER are the following:

- *Start*: when SHIMMER receives this command, it starts collecting temperature measurements in the microSD card. If a Bluetooth connection continues available temperature measurements are also delivered to computer for real-time monitoring and analyses. If no connection is available, temperature measurements are only written in a microSD card. That way, SHIMMER prevents unnecessary use of Bluetooth connection and allows increasing the battery lifetime.
- *Stop*: this command stops the collection of new temperature measurements.
- *Get*: this command performs the transmission of all temperature measurements stored on a microSD card to the computer application for further study and analysis.

Figure 4 presents the diagram of the SHIMMER's operating commands.

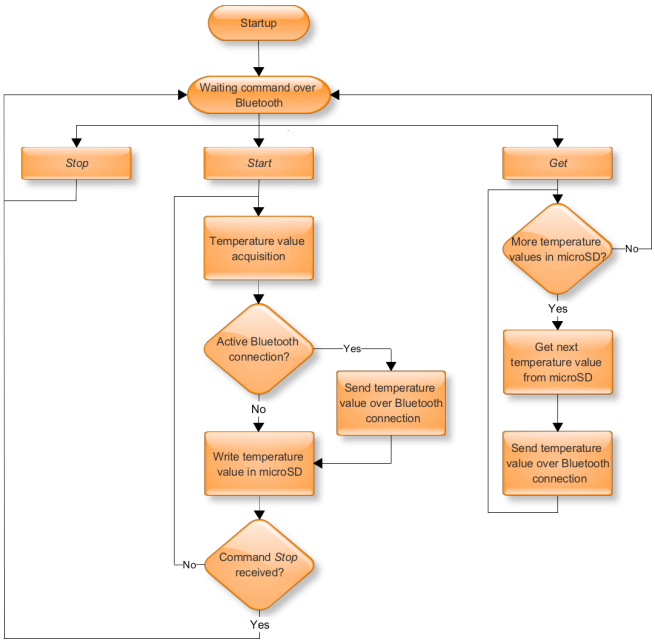


Figure 4. SHIMMER's operating commands diagram.

The analysis of temperature measurements is performed in off-line mode. To ensure good results, as above-mentioned, it is extremely important to know the exact time when each temperature measurement is taken. SHIMMER only has a local time clock, which starts on SHIMMER's start up. To align each measurement with the right global time clock, when a *start* command is sent to SHIMMER, computer also sends its time and date (assuming computer clock is global clock synchronized). This information is then used by SHIMMER's firmware as an offset to local time clock in order to calculate the exact global time clock for every instant of temperature measurements.

V. COMPUTER APPLICATION

A. Application Software Construction

This section describes the created application for collecting, processing, analyzing and visualization the acquired raw data performed by the intra-vaginal sensor in a personal computer. To achieve intra-vaginal sensor temperature values, both real-time and off-line operation modes are available. In the real-time mode, measured values can be achieved when SHIMMER is Bluetooth connected with a computer. In this case, the application software shows the real-time temperature values measured by SHIMMER. Simultaneously, all the temperature measures are sent to the connected computer, via Bluetooth, and at the same time written on the microSD card of SHIMMER platform.

In the off-line mode, the application software can collect all the long-term temperature readings stored in the microSD card. The application software can retrieve all the stored data on the SHIMMER's microSD card. These data may be both visualized and saved in a text file for further use, if needed.

Figure 5 shows de Use Case Diagram of intra-vaginal temperature monitoring application. This use case represents the interaction of the user that can be a physician with this application.

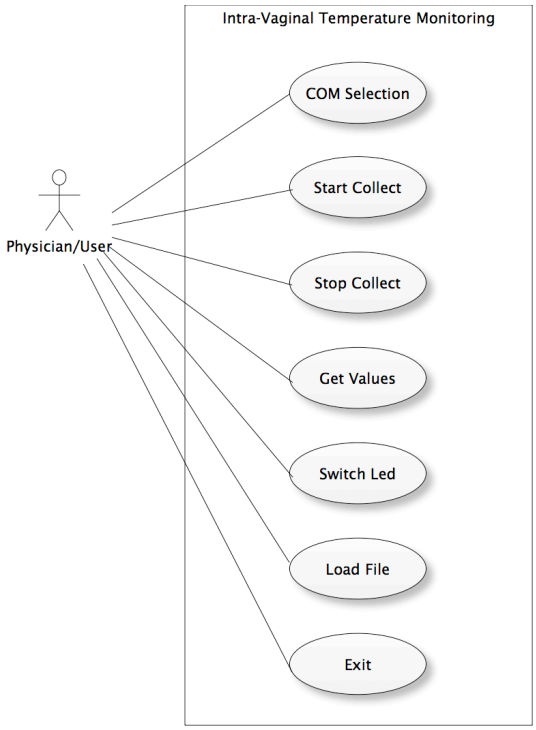


Figure 5. Use case diagram of the intra-vaginal temperature monitoring application.

The description of each use case presented in Figure 5 is the following:

- *COM selection*. This use case allows the physician to select the serial port used for SHIMMER's Bluetooth communication. The selection is performed in a combo box with a list of all the COM ports associated with the SHIMMER platforms paired with the computer. This operation creates two threads. One (*SimpleWriteRead*), is responsible for the communication between SHIMMER and a computer. The other (*GraphTempRT*), represents the real-time temperatures graphic if the intra-vaginal sensor is already in collecting measures mode.
- *Start Collect*. This operation initiates the collection of temperature readings in real-time mode. It creates a *GraphTempRT* thread to represent the real-time graphic temperatures. This operation also

sends to the *SimpleWriteRead* thread a command to be sent through Bluetooth to the sensor's firmware for initialize the temperature measures collection. The temperature readings are then returned from the SHIMMER to the *SimpleWriteRead* thread and are synchronously represented by the *GraphTempRT* thread. The synchronization is needed to allow the integrity of the represented values in a graph – it only goes to the next one if the previous is already presented.

- *Stop Collect*. The *SimpleWriteRead* thread sends the stop collecting command to the SHIMMER over the air. This action also terminates the *GraphTempRT* thread.
- *Get Values*. This action creates a *GraphTempOL* thread to present the graphic representation of the temperature values received from the SHIMMER. A *Get* command is next sent through Bluetooth connection from *SimpleWriteRead* thread to the SHIMMER platform. In return SHIMMER sends all the temperature measures in the microSD to the *SimpleWriteRead* thread. The *GraphTempOL* thread synchronously presents all the measured values retrieved by SHIMMER. Finally, when all values are returned and represented in off-line graph the *GraphTempOL* thread terminates.
- *Switch Led*. This operation is used to confirm if the Bluetooth serial connection between a computer and the SHIMMER platform is working. If the connection is established this operation switches a red LED in SHIMMER. This operation sends the switch LED command from *SimpleWriteRead* thread to the SHIMMER over the air.
- *Load File*. This feature allows an user (physician) to load temperature values from a file. This operation creates a *GraphTempOL* thread used to design the temperature graph of the values read from a file. These values are represented synchronously, and after all the values in the file are presented the *GraphTempOL* thread terminates.
- *Exit*. This action concludes the application execution. Therefore *IVSoftwareJFrame* thread is terminated and consequently the application itself.

Figure 6 presents the class diagram of the computer application for intra-vaginal temperature monitoring.

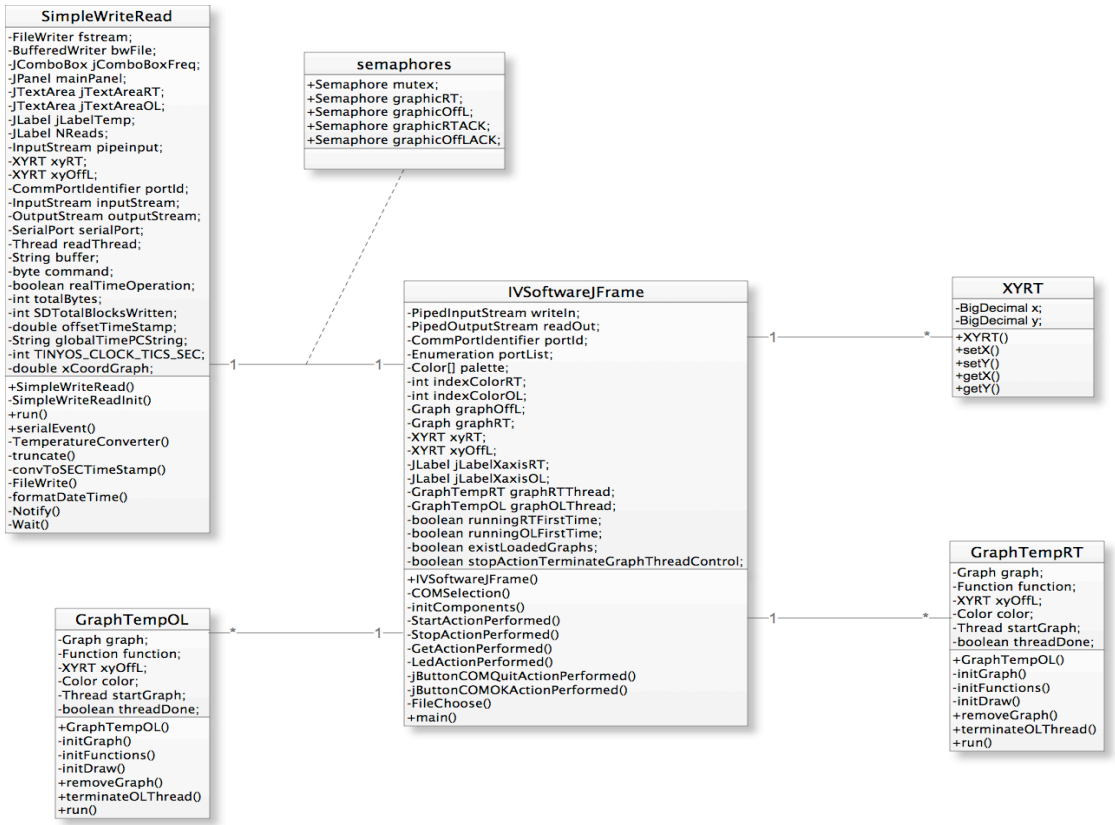


Figure 6. Class diagram of intra-vaginal temperature monitoring application.

B. *Intra-Vaginal Temperature Monitoring Application*

Figure 7 shows the main user's interface of the computer application software. Here, several options are available to configure and interact with SHIMMER platform. This interface also provides a visualization area of graphical temperature readings performed by SHIMMER, in both *real-time* and *off-line* modes.

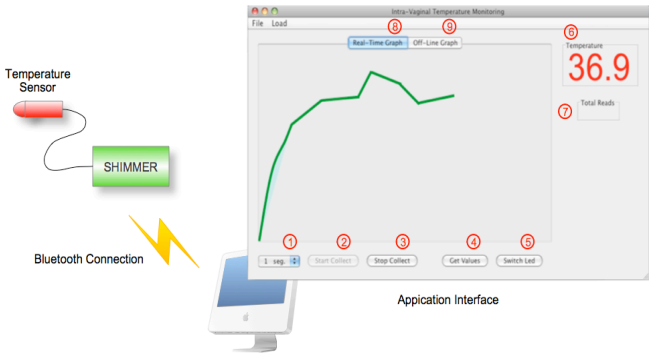


Figure 7. Main window of the intra-vaginal temperature monitoring application.

When the user's interface is open, the *Start Collect* operation is disabled. This behaviour is defined to protect system in such a case that intra-vaginal sensor is already in collection mode. In that mode, if a user starts another collection all previously collected data will be lost. In this case, a *Stop Collect* operation must be firstly performed to guarantee that if sensor is in collection mode no data will be lost accidentally by performing a *Start Collect* operation inadvertently. The intra-vaginal sensor accepts a *Stop Collect* operation *i*) if the sensor is in collected mode, it stops the acquisition of data; and *ii*) if the sensor is in standby mode it is ignored and nothing happening. After performing *Stop Collect* operation the *Start Collect* operation is unblocked.

Below, all the functionalities performed through this window are described in detail.

Definition of Frequency Interval for Measures Acquisition

This feature is performed by the selection of one of the available time values in a combo box identified by ①. This value is used to define the frequency interval of temperature values collected by the sensor. Due to the fact that *Start Collect* operation sends the selected frequency from this combo box to the intra-vaginal sensor, this selection must be performed before each start operation. In acquisition mode, red SHIMMER's LED flashes at each new measure acquisition.

Start Collect operation

The *Start Collect* (②) operation is used to start the acquisition of temperature measurements from the intra-vaginal sensor. This operation also sends to the sensor the current date and time of the computer and the data collection frequency above-described. To perform a *Start Collect* operation, SHIMMER must be in a standby mode. Observing the colour LEDs in this device it can be identified its mode. Green LED *ON* and orange LED flashing indicate standby mode.

Stop Collect operation

Stop Collect (③) operation stops the acquisition of temperature measurements. This operation performs a transition from SHIMMER acquisition mode to standby mode.

Get Values operation

Get Values (④) operation performs the transmission of all stored values in the microSD to the application software. This operation also saves data collected to a text file. This file can be used for persistent storage. When SHIMMER is performed to a *Get Values* operation red LED flashes quickly.

Switch Led operation

Switch Led (⑤) operation is used to verify if SHIMMER is connected to the intra-vaginal application or not. If so, each time a user uses this command, red LED must switch *ON* and *OFF*. Otherwise, if red LED does not switch, this means that SHIMMER is not well recognized by the application. In this case the user should close the application and start again. If the problem persists COM port selected probably is not the one associated with this SHIMMER platform, or SHIMMER could not be well paired with the computer. The user should repeat these configurations and try again the application.

Temperature Information

The label *Temperature* (⑥) is used when SHIMMER is operated in real-time mode and if it is in Bluetooth detection area of a computer. This information field shows the current measured temperature acquired by SHIMMER.

Total Reads Information

The label *Total Reads* (⑦) informs about the number of collected temperature measures stored in the microSD. This information is updated each time *Get Values* operation is performed.

Real-Time Graph

The *Real Time Graph* tab (⑧) is used when a *Start Collect* operation is performed. If SHIMMER is in Bluetooth detection area this field represents a real-time line of temperature measurements. The graph is updated each new measure.

Off-Line Graph

Off-line Graph tab (⑨) presents the temperature graph of all collected temperature measurements stored in the microSD card. This field is filled when a *Get Values* operation is performed.

Load Menu

Load Menu option is used to load saved text files with previous temperature measures into the *Off-Line* tab field. This feature helps on analysis of collected data by the comparison with previous patterns. Various files could be loaded. A different colour line of the graph represents each loaded file.

The proposed application interacts directly with SHIMMER. Then, all the above-described features could only be performed if Bluetooth connection is available for communication between a computer and SHIMMER.

As may be seen in Figure 7, temperature values are presented using a graphical representation. Thus, it is easy to visually identify values outside the normal pattern. These values can lead to sign a range of conditions on female reproductive system (e.g., pregnancy contractions, ovulation period, best fertilization period, etc.). Next, a medical research will carried out with the execution of very important studies to be applied in different kind of gynecology issues, such as, the preterm labor prevention, detection of pregnancy contractions, anticipation and monitoring of the ovulation period (used either as a natural contraception method or, at the opposite, as a estimation of the best fertilization period), effectiveness of some gynecology therapeutics, and supporting the discovery of new possible contraception methods.

VI. SOLUTION VALIDATION

Comfort, usability and, mainly, accuracy were the goals followed in the construction of the intra-vaginal sensor. These targets should be tested and validated in a real environment. Thus, medical team conducted several tests to perform the solution validation.

The biofeedback solution for intra-vaginal temperature monitoring was tested and evaluated in 12 volunteer women. Each woman was monitored for about 60 minutes with temperature readings per second. Simultaneously, women body temperatures under the arm and under the tongue, was measured using a trivial digital thermometer. After analyzing all the temperature data, medical team validated and certified the accuracy of intra-vaginal temperature readings. After concluding the test, each woman was questioned about the possible discomfort caused by the use of the intra-vaginal sensor. None of them shown any issue related to the use of this biosensor.

The sensor was also tested with 8 volunteer women for a longer period on their normal daily life. These tests were preformed in periods of about 2~3 hours. The results were compared to patterns previously defined with traditional thermometers measures in same places above (under the arm and under the tongue). The medical team confirmed the validation of the results.

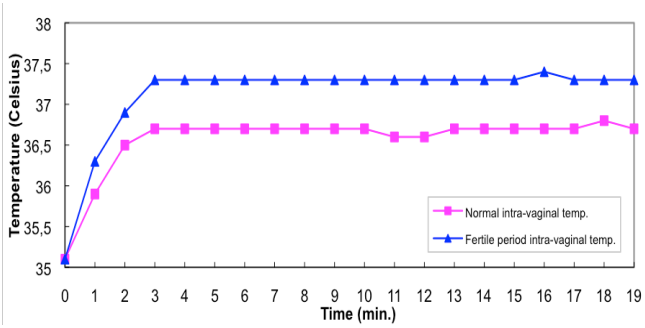


Figure 8. Sample of intra-vaginal temperatures measurements in two different days for the same woman in normal day life.

Figure 8 and Figure 9 present samples of temperature results obtained during tests. In Figure 8 shows temperature values performed by the intra-vaginal sensor for the same woman in tow different days. This graph could be divided in two segments for both curves. First segment, between minute 0 and 2, represents the heating of the sensor to reach ambient temperature (sensor response). The second segment, beyond the 3rd minute, represents the real temperature readings inside vagina. Differences between both temperature readings, beyond the 2nd minute, correspond to the normal intra-vaginal temperature of this woman and her temperature on a fertile period (ovulation period). As expected, intra-vaginal temperature changes according to some female body situations. Monitoring this parameter it could help in medical detection of several situations, such as the preterm labor prevention, detection of pregnancy contractions, anticipation and monitoring of the ovulation period (used either as a natural contraception method or, at the opposite, as a estimation of the best fertilization period), effectiveness of some gynecology therapeutics, and supporting the discovery of new possible contraception methods.

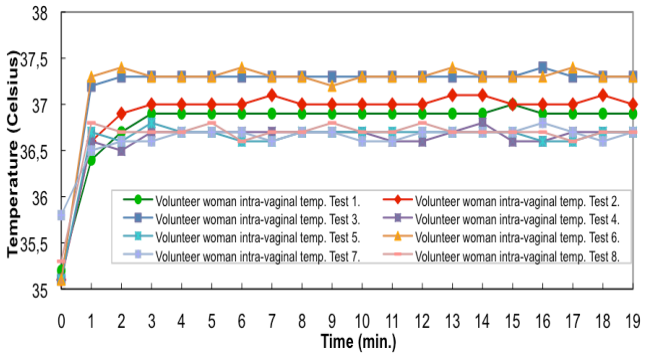


Figure 9. Sample of intra-vaginal temperature measurements for 8 volunteer women.

Figure 9 presents a sample with results of eight tests performed in eight of the twelve volunteers women evaluated. In this graph, the two above-mentioned segments could also be distinguished. From 0 to 2nd minute, it is the response of the sensor and beyond the 3rd minute (after its stabilization), it represents the real temperature values inside women vagina. As may be seen, beyond 3rd minute, differences between both curves means (and confirm) that each woman have her own intra-vaginal temperature. By this reason, individual patterns may have to be established for each woman.

These results represent a great success and encourage medical team for more patients' data collection, using this new biosensor sensor, trying to establish patterns of intra-vaginal temperature behavior. Further, these patterns will be used to understand the relationship between intra-vaginal temperature variations and some reproductive system behaviors, as well as the comparison between patterns and new data collections.

VII. Conclusion and Future Work

The control of women body temperature may help on detection of some symptomatic situations. The environment temperature could influence the skin temperature. This paper introduced a new way to collect women's temperature by the creation of a new biosensor. This biosensor is placed inside women's cervix and collects their core-body temperature. The biosensor can operate in a long-term way, once it has a memory card to store the collected data. The construction of a application software to operate the new biosensor and analyzing the collected data was also described in this paper. The communication mechanisms between the application and the biosensor were also presented. The system was tested, evaluated, and validated by a medical team with a set of 12 volunteer women. The accuracy of the results was also confirmed. Next, medical team will conduct several studies with this system trying to recognize some phenomena related with reproductive system behavior. These studies may contribute for the preterm labor prevention, detection of pregnancy contractions, anticipation and monitoring of the ovulation period, effectiveness of some gynecology therapeutics, and supporting the discovery of new possible contraception methods.

The creation of a miniaturized intra-body biosensor that can be placed inside female cervix, as a hole, instead of placing only the thermistor, may be considered for further developments.

Acknowledgements

Part of this work has been supported by *Instituto de Telecomunicações,* Next Generation Networks and Applications Group (NetGNA), Portugal, in the framework of BodySens Project, by the Euro-NF Network of Excellence of Seven Framework Programme of EU, in the framework of the Specific Joint Research Project PADU, and by Luso-French Program of Integrated University Actions (PAUILF 2010) – Action No. F-TC-10/10.

References

[1] P. Kulkarni and Y. Öztürk, "Requirements and design spaces of mobile medical care," in ACM SIGMOBILE Mobile Computing and Communications Review. vol. 11, pp. 12 - 30, 2007.

[2] G. Shobha, R. R. Chittal, and K. Kumar, "Medical Applications of Wireless Networks," in Proceedings of the Second International Conference on Systems and Networks Communications: IEEE Computer Society, p. 82, 2007.

[3] A. Pantelopoulos and N. Bourbakis, "A survey on wearable biosensor systems for health monitoring," in Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 4887 - 4890, 2008.

[4] C. H. Chan, C. C. Y. Poon, R. C. S. Wong, and Y. T. Zhang, "A Hybrid Body Sensor Network for Continuous and Long-term Measurement of Arterial Blood Pressure," in International Summer School and Symposium on Medical Devices and Biosensors: 4th IEEE/EMBS, pp. 121 - 123, 2007.

[5] S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. H. Growdon, M. Welsh, and P. Bonato, "Analysis of Feature Space for Monitoring Persons with Parkinson's Disease With Application to a Wireless Wearable Sensor System," in Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual Int. Conference of the IEEE, pp. 6290-6293, 2007.

[6] W. D. Jones, "Taking Body Temperature, Inside Out," in IEEE Spectrum. vol. 43 Issue 1, pp. 13-15, 2006.

[7] U. Anliker, J. A. Ward, P. Lukowicz, G. Troster, F. Dolveck, M. Baer, F. Keita, E. B. Schenker, F. Catarsi, L. Coluccini, A. Belardinelli, D. Shklarski, M. Alon, E. Hirt, R. Schmid, and M. Vuskovic, " AMON: a wearable multiparameter medical monitoring and alert system," in IEEE Transactions on Information Technology in Biomedicine vol. 8, Issue 4, pp. 415 - 427, 2004.

[8] H. W. Taylor, S. E. Shidler, B. L. Lasley, L. Ngalamou, and F. E. Taylor, "FSH biosensor to detect postpartum ovarian recrudescence," in Engineering in Medicine and Biology Society. IEMBS '04. 26th

Annual International Conference of the IEEE. vol. 1, pp. 1998 - 2001, 2004.

[9] F. Nebeker, "Golden accomplishments in biomedical engineering," in Engineering in Medicine and Biology Magazine, IEEE. vol. 21, pp. 17-47, 2002.

[10] J. Beaudoin and R. Marrocco, "Attentional validity effect across the human menstrual cycle varies with basal temperature changes," in Behavioural brain research. vol. 158, pp. 23-29, 2005.

[11] I. Campbella, "Body temperature and its regulation," in Anaesthesia & Intensive Care Medicine. vol. 9, pp. 259-263, 2008.

[12] E. F. J. Ring, "Progress in the measurement of human body temperature," in Engineering in Medicine and Biology Magazine, IEEE. vol. 17, pp. 19-24, 1998.

[13] L. Ngalamou and D. Rose, "Fertility information appliance," in Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002). pp. 335- 338, 2002.

[14] Z. McCreesh and N. Evans, "Radio telemetry of vaginal temperature," in Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE, Baltimore, MD, Nov. 03-06, vol. 2, pp. 904-905, 1994.

[15] Z. McCreeshab, N. E. Evans, and W. G. Scanlonab, "Vaginal temperature sensing using UHF radio telemetry," Medical Engineering & Physics Journal by Elsevier Inc., vol. 18, pp. 110-114, 1996.

[16] W. Jones, "Taking Body Temperature, Inside Out," IEEE Spectrum online, pp. 13-15, January 2006.

[17] DuoFertility, "http://www.duofertility.com", accessed in Feb. 2009.

[18] Z. McGreesh, N.E. Evans, and W.G. Scanlon, "Vaginal temperature sensing using UHF radio telemetry", Journal of Medical Engineering and Physics, Vol. 18, No. 2, pp. 110-114, March 1996

[19] D. H. Kosted, "A Method and System of Continual Temperature Monitoring", US Patent 20070027403, Feb 2007, http://www.freepatentsonline.com/US20070027403.html

[20] M. H. James and T. G. Knowles, "Method of detecting and predicting ovulation and the period of fertility," WO/2008/029130, 13/03/2008.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601