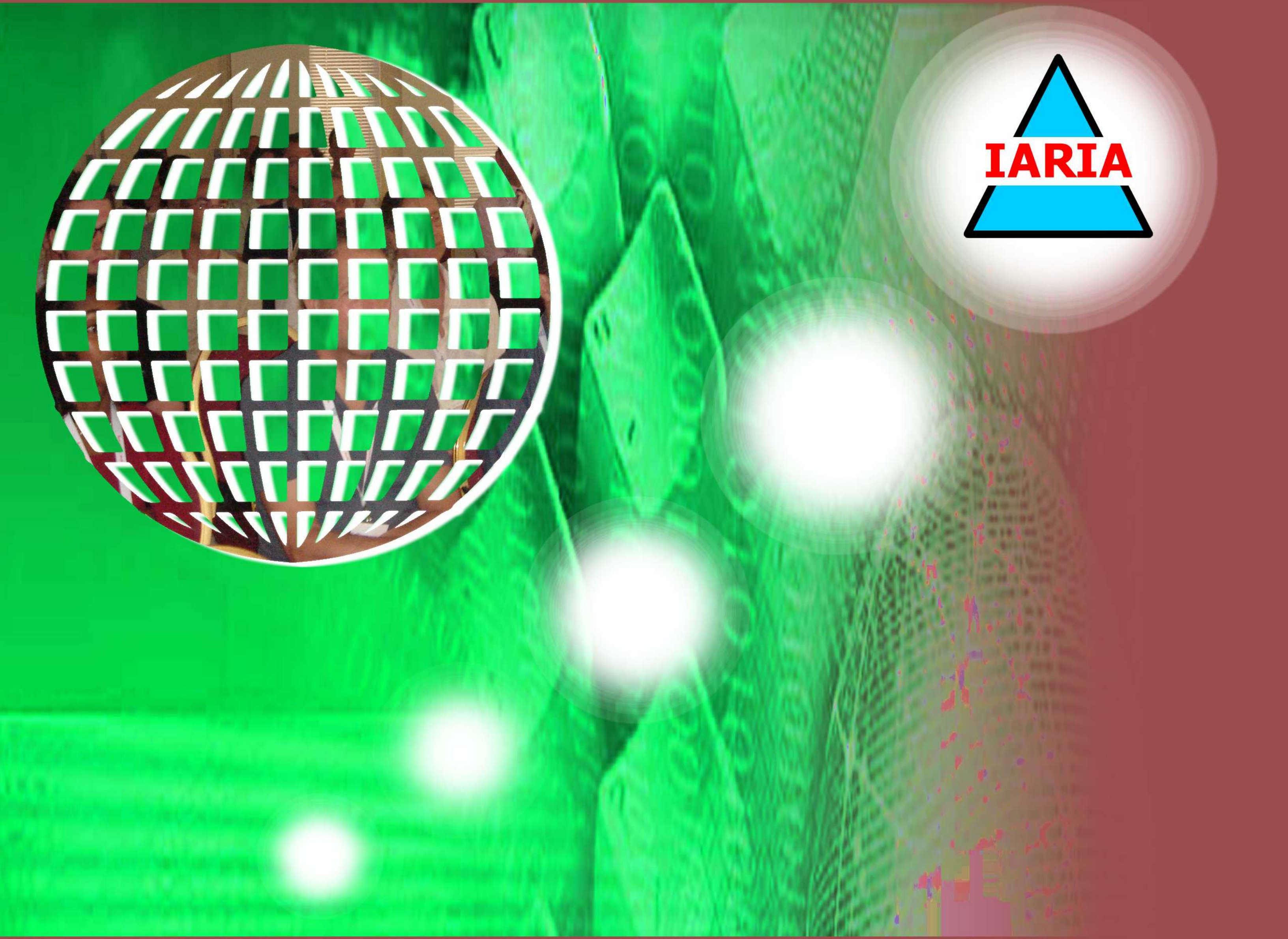
International Journal on Advances in Life Sciences



2022 vol. 14 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 14, Number 3 & 4, 2022

Udi Davidovich, Amsterdam Health Service - GGD Amsterdam, The Netherlands
Maria do Carmo Barros de Melo, Telehealth Center, School of Medicine - Universidade Federal de Minas Gerais (Federal University of Minas Gerais), Brazil
Kari Dyb, Norwegian Centre for Integrated Care and Telemedicine / University Hospital of North Norway | University of Tromsø, Norway
Juergen Eils, DKFZ, German
Anne G. Ekeland, Norwegian Centre for Integrated Care and Telemedicine / University Hospital of North Norway | University of Tromsø, Norway
El-Sayed M. El-Horbaty, Ain Shams University, Egypt
Ivan Evgeniev, TU Sofia, Bulgaria
Karla Felix Navarro, University of Technology, Sydney, Australia
Joseph Finkelstein, The Johns Hopkins Medical Institutions, USA
Stanley M. Finkelstein, University of Minnesota - Minneapolis, USA
Adam M. Gadomski, Università degli Studi di Roma La Sapienza, Italy
Ivan Ganchev, University of Limerick, Ireland / University of Plovdiv “Paisii Hilendarski”, Bulgaria
Jerekias Gandure, University of Botswana, Botswana
Xiaohong Wang Gao, Middlesex University - London, UK
Josean Garrués-Irurzun, University of Granada, Spain
Hassan Ghazal, Moroccan Society for Telemedicine and eHealth, Morocco
Piero Giacomelli, SPAC SPA -Arzignano (Vicenza), Italia
Alejandro Giorgetti, University of Verona, Italy
Anthony Glascock, Drexel University, USA
Wojciech Glinkowski, Polish Telemedicine Society / Center of Excellence "TeleOrto", Poland
Francisco J. Grajales III, eHealth Strategy Office / University of British Columbia, Canada
Conceição Granja, Conceição Granja, University Hospital of North Norway / Norwegian Centre for Integrated Care and Telemedicine, Norway
William I. Grosky, University of Michigan-Dearborn, USA
Richard Gunstone, Bournemouth University, UK
Amir Hajjam-El-Hassani, University of Technology of Belfort-Montbéliard, France
Lynne Hall, University of Sunderland, UK
Päivi Hämäläinen, National Institute for Health and Welfare, Finland
Anja Henner, Oulu University of Applied Sciences, Finland
Stefan Hey, Karlsruhe Institute of Technology (KIT) , Germany
Dragan Ivetic, University of Novi Sad, Serbia
Sundaresan Jayaraman, Georgia Institute of Technology - Atlanta, USA
Malina Jordanova, Space Research & Technology Institute, Bulgarian Academy of Sciences, Bulgaria
Attila Kertesz-Farkas, University of Washington, USA
Hassan Khachfe, Lebanese International University, Lebanon
Valentinas Klevas, Kaunas University of Technology / Lithuaniain Energy Institute, Lithuania
Anant R Koppar, PET Research Center / KTwo technology Solutions, India
Bernd Krämer, FernUniversität in Hagen, Germany
Ramesh Krishnamurthy, Health Systems and Innovation Cluster, World Health Organization - Geneva, Switzerland
Roger Mailler, University of Tulsa, USA
Dirk Malzahn, OrgaTech GmbH / Hamburg Open University, Germany
Salah H. Mandil, eStrategies & eHealth for WHO and ITU - Geneva, Switzerland
Herwig Mannaert, University of Antwerp, Belgium
Agostino Marengo, University of Bari, Italy
Igor V. Maslov, EvoCo, Inc., Japan
Ali Masoudi-Nejad, University of Tehran , Iran
Cezary Mazurek, Poznan Supercomputing and Networking Center, Poland
Teresa Meneu, Univ. Politécnica de Valencia, Spain
Kalogiannakis Michail, University of Crete, Greece
José Manuel Molina López, Universidad Carlos III de Madrid, Spain

International Journal on Advances in Life Sciences
Volume 14, Numbers 3 & 4, 2022

CONTENTS

Sensor Glove Approach for Continuous Recognition of Japanese Fingerspelling in Daily Life

Yuhki Shiraishi* Akihisa Shitara†, Fumio Yoneyama* and Nobuko Kato*
*Faculty of Industrial Technology, Tsukuba University of Technology, Japan
Email: {yuhkis, yonefumi, nobuko}@a.tsukuba-tech.ac.jp
†Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan
Email: theta-akihisa@digitalnature.slis.tsukuba.ac.jp

Abstract—**To achieve smooth communication between the deaf and hard of hearing and hearing people, we developed a Japanese fingerspelling (JF) recognition system based on sensor gloves. A light and inexpensive sensor glove was adapted for the daily use of the system. We conducted evaluation experiments using a convolutional neural network (CNN) to recognize 76 characters in JF. The target JF alphabet included 35 characters for dynamic fingerspelling, and required both finger and wrist movement. The experimental results show that the average recognition rate of the developed system was approximately 70.0%. Additionally, we conducted a continuous fingerspelling recognition experiment using CNNs and long short-term memory (LSTMs) networks, aiming to recognize consecutive fingerspelling. We proposed a dataset to exploit the characteristics of JF and selected 64 words according to the finger flexion, direction, and movement differences among various signers. Using the collected data, we then conducted evaluation experiments with seven types of neural networks. The overlapping characteristics present in JF were exploited because finger flexion, finger extension, hand direction, and hand movements vary significantly among people currently learning sign language, people corresponding in Japanese sign language (JSL), and people using JSL in their daily lives. Consequently, the average recognition rate (micro F-measure) of 76 JF characters was approximately 92.1%. Based on the results of single fingerspelling and continuous fingerspelling recognition experiments, we discussed the issues concerning the recognition of JF characters and development of sign language recognition systems.**

Keywords–Sign language; Japanese fingerspelling; Sensor glove; Recognition; Convolutional neural network; Long short-term memory.

I. Introduction

In this study, we developed a Japanese fingerspelling (JF) recognition system incorporating a sensor glove and deep learning to achieve smooth communication between the deaf and hard of hearing (DHH) and hearing people, and investigated the recognition rate of the JF alphabet.

This study extends our initial study [1] on a sensor-glove-based JF recognition system for using deep learning to realize smooth communication between DHH and hearing people.

In recent years, interest in speech recognition and information technology devices with voice input functions has increased. Various applications, such as UDtalk [2], KoeTra [3], and cloud-speech-to-text services [4] have been released to provide information accessibility to the DHH based on speech recognition. Consequently, the DHH can read text corresponding to the speech of hearing people.

As a primary communication method, sign language is used in everyday conversations among the DHH. However, hearing people find reading sign language difficult; this results in a communication gap between DHH and hearing people.

Sign language has different characteristics from spoken language. It is expressed itself through finger extensions and flexion, hand directions, hand movements, and facial expressions. Hence, learning and reading sign language is difficult. Therefore, a system for converting sign language into voice information or text information (i.e., a sign language recognition system) is necessary (see Figure 1).

Research has been conducted on information accessibility systems for sign language recognition [5]–[12]. However, compared with information accessibility systems based on speech recognition that are fast reaching maturity, the development of a practical sign language recognition system remains in progress.

In this context, even in a specific country, for example, Japan, differences exist in sign language expressions in the daily lives of people new to sign language, those using signed exact Japanese (SEJ), and those using Japanese Sign Language (JSL). A person learning sign language for the first time learns sign language using a dictionary and other teaching materials. Sign language dictionaries contain many standardized finger expressions for people to imitate and practice, and for slowly and carefully expressing themselves. People using SEJ express themselves one word at a time, as in Japanese, and not using facial expressions. Conversely, people using JSL express themselves using their fingers, hand directions, hand movements, facial expressions, etc. People with relatively little experience in sign language tend to express themselves slowly, whereas people with more experience tend to express themselves quickly.

In addition to using sign language, the DHH people use fingerspelling, e.g., to express their names, proper nouns, and words not present in JSL. As mentioned earlier, finger flexion, hand directions, and hand movements can vary depending on if the person is new to sign language, uses SEJ, or uses a JSL. For example, although the finger positions of “ka” and “ga” are identical, the hand movements are different. In this case, an evident difference exists between the hand movements among the three groups, i.e., those new to the sign language, those using an SEJ, and those using a JSL.

In this study, we developed a JF recognition system based on a sensor glove, deep learning, and acquired data on finger flexion, hand directions, and hand movements for JF signing used in daily life.

A sign language recognition system must recognize hand positions, directions, shapes, and motions. Methods for recognizing sign language can be classified broadly into non-contact

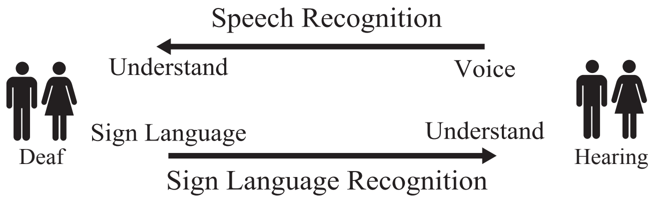


Figure 1. Information accessibility system.

TABLE I. NUMBER OF FINGERSPELLING CHARACTERS IN DIFFERENT COUNTRIES.

Language	*Dynamic*	*Static*	*Sum*
American	2	24	26
French	3	23	26
Japanese	35	41	76

approaches such as recognition using cameras [5] [6] [10], and contact approaches, such as those using sensor gloves [7] [8] [11] [12].

Luzhnica et al. [7] reported a recognition accuracy of 98.5% for sign language using a sensor glove; however, they only considered approximately 30 recognition candidate classes, making this method insufficient for practical use.

In recent years, technologies based on deep learning have attracted significant attention. By increasing the number of hidden layers in a neural network, we can improve the recognition rates of deep learning, which is a type of machine learning. Various techniques for applying deep learning have been reported for improving the gesture recognition accuracy based on image recognition [5].

A camera is a non-contact-type sensor, but is difficult to use for sign language recognition in daily life because it is easily affected by environmental factors, complicating its use in different environments. In addition, when standing in front of people and a camera at a lecture, a speaker tends to speak to the camera without looking at the people as necessary, thereby constraining the speaker from making a connection with the audience. In contrast, hand shape recognition using contact sensors such as sensor gloves is easier, because the sensors are attached directly to the hands.

We were motivated by the goal of improving recognition accuracy using conductive fiber weaving technology [13], as this technology can reduce the weight and cost of sensor gloves and simplify hand movements used in daily life for easy recognition by deep learning (see Figure 2).

In our experiments, we evaluated our developed system by classifying 76 characters of the JF alphabet, including dynamic (non-static) fingerspelling characters; those are a unique feature of JF compared to other fingerspelling systems, as shown in Table I.

The evaluation experiments for a single JF were conducted using a convolutional neural network (CNN) as a learning model (this type of model performed the best in previous studies) to reduce the data reduction by calculating the moving averages of the data acquired from gyro sensors. In these experiments, all 76 JF characters of JF were included as recognition targets, as were dullness, semi-voiced sounds, diphthongs, and long vowels. These experiments were conducted using all the collected data under various experimental conditions.

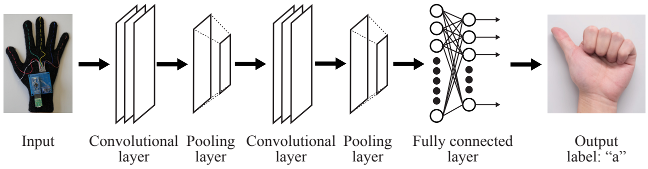


Figure 2. Recognition diagram.

In the continuous JF recognition evaluation experiment, we utilized the neural network constructed for the single JF recognition task as a learning model, introduced a long short-term memory (LSTM), and built seven types of neural networks. As in the evaluation experiment for single-finger character recognition, all 76 characters of JF are used for recognition. Furthermore, evaluation experiments were conducted for consecutive finger characters with two or more characters. In this experiment, we propose a dataset that exploits the characteristics of JF, and selected 64 words owing to the differences in finger flexion, directions, and movements differences among people new to sign language, people using SEJ, and people using JSL. We then conducted evaluation experiments with the seven types of neural networks using the collected data.

This study provides the following contributions:

- the development and evaluation of a fingerspelling recognition system using an inexpensive and lightweight sensor glove;
- the development and evaluation of a continuous JF recognition system using CNNs and LSTMs; and
- the proposal and evaluation of a dataset for fingerspelling recognition in the daily lives of various signers.

In Section II, we introduce the related research results. In Section III, we describe the single fingerspelling recognition experiments. In Section IV, we describe the continuous fingerspelling recognition experiments. In Section V, we provide our conclusions.

II. RELATED WORK

Previous research on fingerspelling recognition has proposed two types of sensors for recognizing a series of operations in fingerspelling: contact-type sensor gloves and non-contact-type cameras.

A. *Image recognition*

Several methods have been proposed for recognizing hand shapes based on processing images of fingerspelling as captured by cameras. Mukai et al. [9] reported that a fingerspelling recognition method targeting 41 immobile characters in JSL resulted in an average recognition accuracy of 86%. They used a classification tree and machine learning based on a support vector machine to classify individual images. Hosoe et al. [10] used deep learning for recognition and achieved a recognition rate of 93%, but only for static fingerspelling. Jalal et al. [6] reported a recognition rate of 99% for American sign language (ASL) images based on a deep learning algorithm for static fingerspelling (i.e., excluding "J" and "Z"). However, the recognition accuracy could not be considered as sufficient for practical recognition in JF. Additionally, relatively few recognition results have been reported for dynamic fingerspelling (i.e., fingers moving when expressing a character). In a study of dynamic fingerspelling in JSL [14], the identification of hand shapes was performed using a kernel orthogonal mutual subspace method from images of hand regions obtained from distance images, and the classification of movements was performed using decision trees based on center-of-gravity coordinates. These results yielded a 93.8% identification rate. However, the recognition accuracy was insufficient for the practical recognition required for JF.

B. Sensor glove recognition

Several methods have been proposed for recognizing hand shapes based on measurement data acquired by contact-type sensor gloves. These methods can measure finger flexion, hand positions, and directional data. The measurement data are then sent to a personal computer, and a classification algorithm is used to recognize hand shapes. Cabrera et al. [11] paired the Data Glove 5 Ultra [15] sensor glove with an acceleration sensor to acquire information regarding the degree of flexion of each finger and wrist direction. They conducted test classification using 24 static fingerspelling characters in ASL, excluding "J" and "Z." Their neural network was trained using 5 300 patterns and achieved a recognition rate of 94.07% for 1 200 test patterns. Mummadi et al. [12] prototyped a sensor glove with multiple embedded inertial sensors. They collected French sign language fingerspelling data from 57 people and achieved an average recognition rate of 92% with an F1-score of 91%. Kakoty et al. [16] reported on a dataset of one-handed Indian sign language alphabets (C, I, J, L, O, U, Y, W), ASL alphabets (A to Z), and signed numbers (0 to 9), using a radial basis function with 10-fold cross-validation Using a kernel-supported vector machine, they achieved an average recognition rate of 96.7% and reported that the data were converted to speech. Chong et al. [17] placed six inertial measurement units (IMUs) on the back of the palm and on each fingertip to capture their motion and orientations. Ultimately, 28 proposed word-based sentences in ASL were collected, and 156 features were extracted from the collected data for classification. Using the long short-term memory (LSTM) algorithm, the system achieved an accuracy of up to 99.89%. Notably, 12 people cooperated with us in the data collection experiment, but whether they were deaf or hearing people was unclear. Yu et al. [18] reported on the architecture of a data glove system comprising a stnm32MCU, flex4.5 bending sensor, mpu6050 six axis sensor, Bluetooth transmission module, and cellphone voice application. The system was developed and connected to a Java-based processing software. They reported that their system recognized sign language movements and could output the words to be said using the intelligent voice system. However, the glove does not feature global movement and rotation tracking. Glauser et al. [19] demonstrated a glove's performance in a series of ablation experiments while exploring various models and calibration methods. However, the glove does not come with a global translation and rotation tracking. Realizing a sign language recognition system requires hand orientations and motions. Among the various methods for performing JF recognition, the conductive fiber braid method [13] uses gloves woven with conductive fibers instead of flexion sensors. These gloves can recognize hand shapes and movements as they are directional gyro sensors incorporated into them. However, the recognition rate for JF ("a," "i," "u," "e," "o") based on Euclidean distance has been reported as only 60%.

C. Data collection

Regarding image recognition, several large-scale continuous sign language recognition (CSLR) benchmarks have been published [20]. For example, we introduced three large-scale CSLR benchmarks: PHOENIX-2014, Chinese sing language (CSL), and PHOENIX-2014-T. PHOENIX-2014 is a publicly available German Sign Language dataset and the most famous CSLR benchmark. This corpus is taken from broadcast news regarding the weather. The CSL dataset consists of 100 sign language sentences and 178 words related to everyday life. Fifty signers performed each sentence, resulting in 5,000 videos in total. A matched isolated CSL database containing 500 words is also provided for pre-learning. Each word was performed 10 times by 50 signers. PHOENIX-2014-T annotates the new videos with two annotations: the sign language terms for the CSLR task, and the German translation for the a sign language translation (SLT) task. The vocabulary consists of 1,115 terms for sign language and 3,000 for German. This dataset is available in [21]. However, the data of these three large-scale CSLR benchmarks are insufficient to realize a highly accurate sign language recognition system using deep learning. Further research is being conducted to increase the amount of available data.

Extensive data for image recognition can be obtained from online sources. For example, the Shi et al. [22] dataset contains clips of fingerspelling sequences cut from sign language "in the wild" videos obtained from online sources such as YouTube and dafvideo.tv [23]. The datasets contain 5,455 training sequences from 87 signers of "ChicagoFSWild," 981 development (validation) sequences from 37 signers, and 868 test sequences from 36 signers, without overlapping signers among the three sets. Another dataset, "ChicagoFSWild+," contains 50,402 training sequences from 216 signers, 3115 development sequences from 22 signers, and 1,715 test sequences from 22 signers. Compared to ChicagoFSwild, the crowdsourcing setup of ChicagoFSWild+ enables the collection of considerably more training data while significantly reducing the efforts of experts and researchers.

Danielle et al. [24] expressed privacy concerns regarding contributing to a filtered sign language corpus, using very expressive avatars and blurred faces, which may affect the willingness to participate. Training on filtered data may improve the recognition accuracy. In the case of camera recognition, the look of the face is also captured; thus, privacy must also be considered. In contrast, sensor glove recognition does not require pictures of the face; thus privacy concerns are reduced and the data can be more simply collected.

III. SINGLE FINGERSPELLING RECOGNITION EXPERIMENT

To achieve smooth communication in real-world environments, we designed a system for communicating information using lightweight and comfortable sensor gloves for recognizing fingerspelling with high accuracy in real time. The

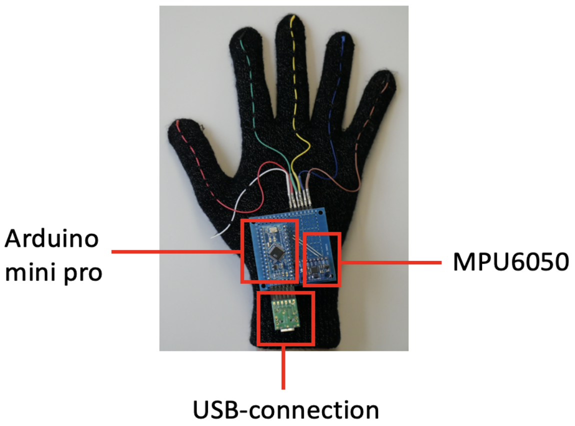


Figure 3. Prototype of sensor glove.

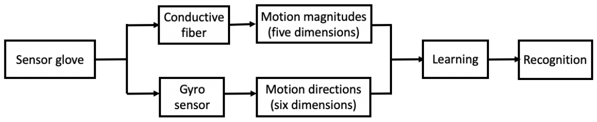

Sensor glove
Conductive fiber
Gyro sensor
Motion magnitudes (five dimensions)
Motion directions (six dimensions)
Learning
Recognition

Figure 4. Software structure.

developed system consists of a sensor value measurement unit and recognition unit. Figure 3 shows the JF recognition system developed in this study. Figure 4 shows the corresponding software architecture.

A. *Sensor glove*

To efficiently recognize fingerspelling efficiently based on hand, finger, and wrist data, detecting motion magnitudes and directions using the sensor glove is necessary. In this study, we adopted a hand shape recognition technique using conductive fiber sensor gloves, which are more comfortable, less expensive, and lighter than traditional sensor gloves. Motion directions are detected using a gyro sensor, whereas motion magnitudes are detected based on resistance changes in the conductive fibers of the gloves. The motion detection board is an Arduino board and the measurement values from the sensor glove are transferred from the detection board to a PC, where they are saved in comma-separated-value format. The machine learning and motion recognition are performed using Python implementations on a PC. The sensor readings for JF motion from the data gloves have different scales depending on the wearer. Therefore, the data are subjected to linear normalization in consideration of the differences in movement. Additionally, because the activation and likelihood functions of the proposed system are based on probabilities, as a prepossessing for the network inputs, we perform scale conversion to a range of zero to one.

The motion magnitudes are detected based on the resistance changes in the conductive fibers during flexion and extension of the fingers. We use partial pressure values to calculate the input voltages based on (1).

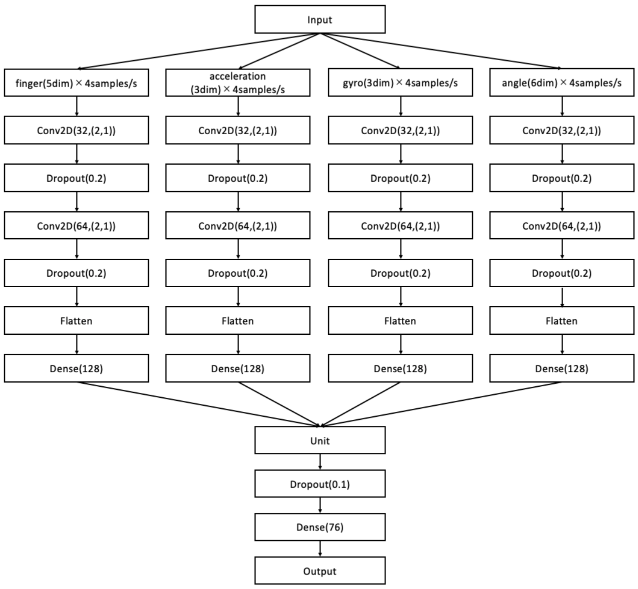


Figure 5. Architecture of the convolutional neural network.

$$V_{in} = \frac{R_1}{R_1 + R_2} * V_{out} \tag{1}$$

In this equation, V_{in} is the estimated motion magnitude, V_{out} is the reference voltage, R_1 is the variable resistance of the conductive fibers, and R_2 is a fixed resistance. When a finger is stretched, the resistance value of the conductive fiber increases. When a finger is bowed, the resistance value of the fiber decreases.

B. *Recognition algorithm*

In this study, we adopted a CNN. This type of network has achieved high recognition rates in previous studies. The CNN and k-fold cross-validation are implemented using the open-source libraries, TensorFlow [25] and scikit-learn [26]. We also adopted the RMSprop training algorithm [27]. The activation function is a rectified linear unit, as shown in (2). The error function is the cross-entropy function shown in (3), where t_k is the correct label (one-hot expression) and y_k expresses the network output.

$$f(u) = max(u, 0) \tag{2}$$

$$E = -\sum_k t_k \log y_k \tag{3}$$

The main features of CNNs are the convolutional and pooling layers. These layers are updated as their feature values are extracted during the training process. We transform the measurement data acquired by the sensor glove into two dimensions based on training and evaluation trials. The motion magnitudes, accelerations, and gyro readings are branched at the time of input. Through the CNN (typical layer size of 32 to 64 nodes), these data are coupled using “Flatten” and “Dense” operations (128 nodes). Finally, the outputs are generated using

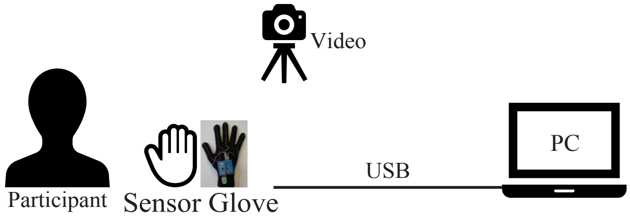


Figure 6. Data acquisition experiment.

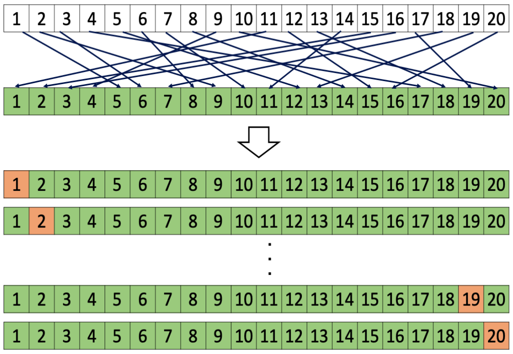


Figure 7. Twenty-fold cross-validation by shuffling data.

TABLE II. TWENTY-FOLD CROSS-VALIDATION RESULTS.

k	Learning data (%)	Validation data (%)
1	93.6	65.0
2	94.1	75.5
3	94.8	68.7
4	93.1	69.7
5	94.2	66.3
6	93.9	73.2
7	92.9	67.9
8	93.5	71.1
9	93.0	67.4
10	94.6	70.5
11	93.4	71.6
12	93.0	66.1
13	94.6	68.9
14	94.3	70.3
15	93.0	69.7
16	93.4	68.4
17	92.9	71.3
18	93.1	71.1
19	94.5	74.2
20	94.5	72.4
Average	93.7	70.0

TABLE III. MISRECOGNITION PATTERNS.

Teacher	a	sa	ku	yo	ke	te	ki	chi	chi
Prediction	sa	a	yo	ku	te	ke	chi	ki	tsu
Rate (%)	21.0	19.0	14.0	20.0	12.0	28.0	12.0	12.0	34.0
Teacher	tsu	ni	ha	ne	ma	hi	re	wo	xya
Prediction	chi	ha	ni	ma	ne	re	hi	xya	wo
Rate (%)	32.0	20.0	22.0	13.0	11.0	19.0	23.0	11.0	13.0
Teacher	gi	di	ge	de	di	du	zo	bu	
Prediction	di	gi	de	ge	du	di	bu	zo	
Rate (%)	12.0	13.0	29.0	20.0	39.0	35.0	14.0	15.0	

an additional Dense operation (76 nodes) corresponding to the number of JF characters, outputs are generated. Figure 5 presents a system overview of the CNN. In the CNN, inputs are initially separated based on the physical meanings of each signal. The separated signals are eventually combined to recognize JF characters.

C. Data collection

To target the 76 JF characters, we recruited 20 participants (from 20 to 27 years old). In our experiments, each participant wore a sensor glove and performed the motions of the fingerspelling characters in sequence for 1 s at a time according to directions provided by a moderator. As shown in Figure 6, video was also recorded to capture the motions of the wrists and fingers of the participants. For each 1 s motion and at a rate of 200 samples per second (sps), the sensor gloves captured five dimensions of motion magnitude data, three dimensions of acceleration data, and three dimensions of gyro data, to obtain data for 11 dimensions. Data labeling was conducted manually and simultaneously with the data collection. This series of motions was repeated five times. Therefore, with five repetitions per participant, 76 JF characters, and 200 sps for 1 s, a total of 76,000 motion measurement data were collected for each participant. We were able to collect a total of 1,520,000 data samples from all 20 participants. These experiments were conducted with approval from the Tsukuba University of Technology Research Ethics Committee (approval number: H30-17).

First, we performed extensive data cleaning and feature selection operations. T o prevent gyro drift, we used Madgwick filters [28] to calculate angles from the values of the acceleration and gyro sensors in real time. This enabled calculations of three angle dimensions from the acceleration and gyro data. To clarify the hand directions, the angles were converted into sine and cosine data. The resulting six dimensions were combined with the aforementioned motion magnitudes (five dimensions) and motion directions (six dimensions) mentioned above to generate a total of 17 dimensions. Next, we conducted a review of the sampling frequency. Although 200 sps could be acquired without leakage, noise and training times were included in these samples. Therefore, the number of data was reduced by calculating a moving average to achieve a final value of 4 sps.

D. Evaluation experiments

The collected data were evaluated using the CNN (Figure 5) and k-fold cross-validation (k = 20). In our evaluation experiments, data shuffling was performed using Google Colaboratory [29]. The number of folds for the k-fold cross-validation was set to 20 according to the number of participants. Additionally, confusion matrices and accuracy rates were generated using 20-fold cross-validation for all data shuffling evaluations (see Figure 7).

E. Results and discussion

The experimental results from the 20-fold cross-validation are listed in Table II. This table reveals an average recognition rate of approximately 70.0%.

As shown in Figure 8 and Table III, various misrecognition patterns occurred. We believe these patterns occurred because

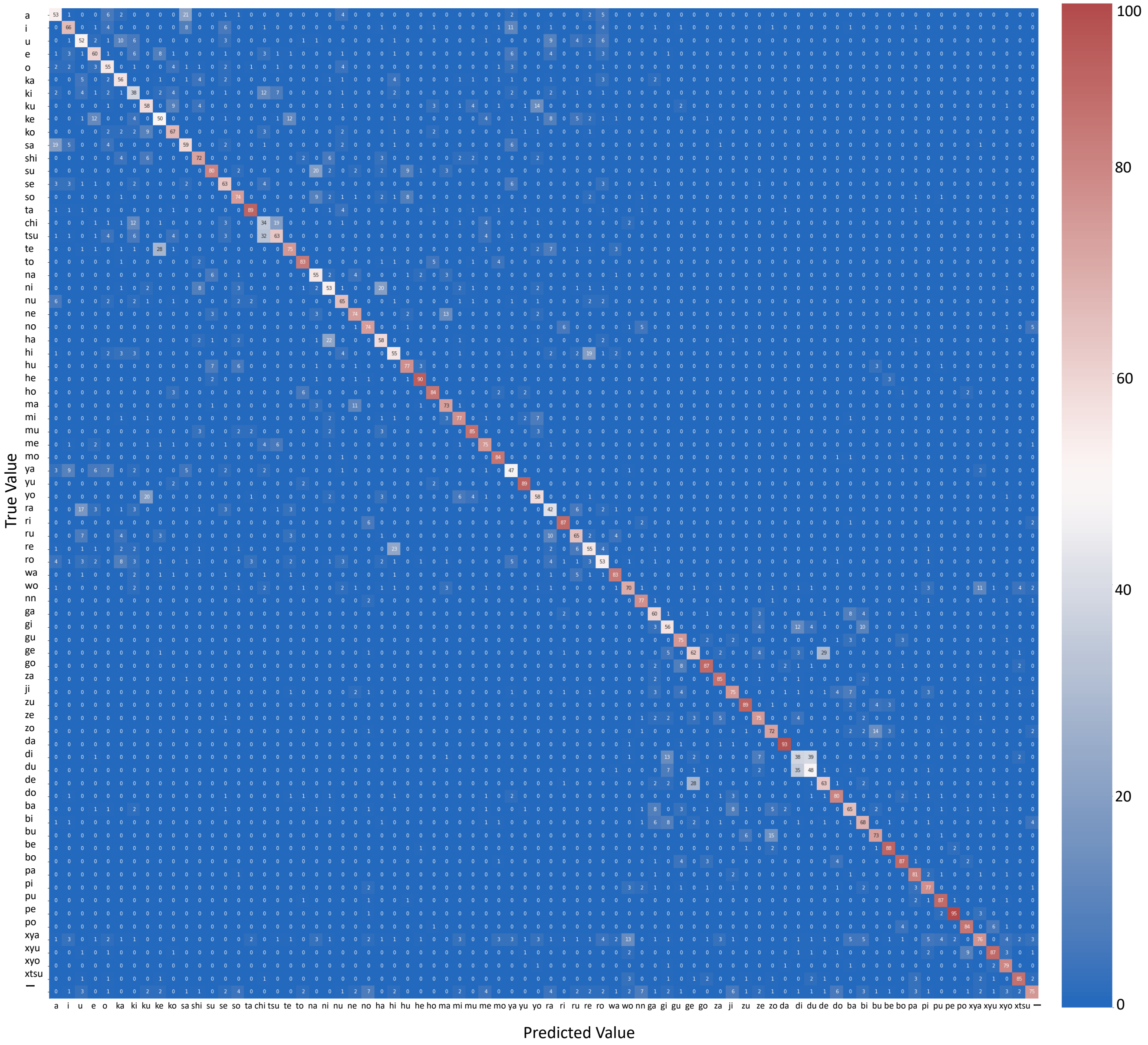


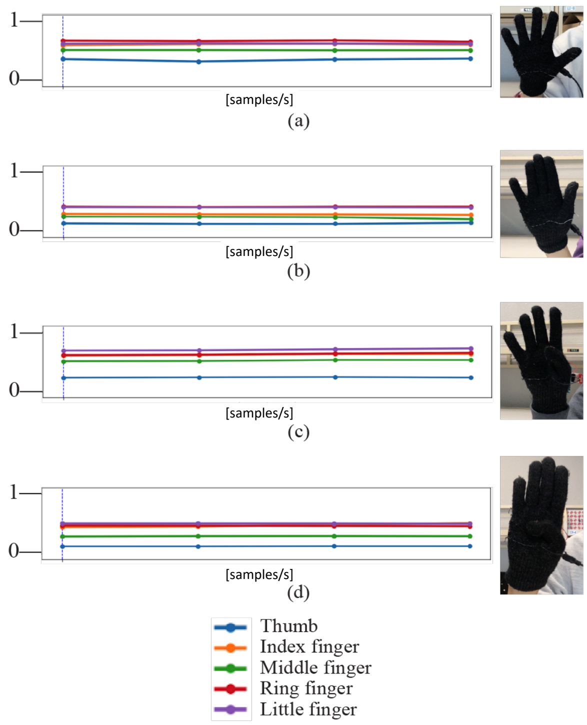
Figure 8. Confusion matrix.

the conductive fibers are firmly attached to the sensor gloves. We confirmed that the hand directions for “ha” and “ni”, which are JF characters, varied among participants. Additionally, “ne” and “ma” appear to be confused based on both hand bending and finger bending.

Figure 9 presents the sample input data leading to misrecognition for the JF characters “te” and “ke”. By analyzing the data, it was confirmed that the close contact between the fingers caused these errors. Notably, the thumb sometimes contacted the forefinger. Additionally, depending on the participant, the hand could be widely opened or the fingers could be in close contact.

Figure 10 presents examples of acquiring data from two participants using the sensor glove for dynamic fingerspelling. This figure clearly highlights the individual differences in fingerspelling between the participants, particularly in the strength of the finger bending (including noisy signals), timing of hand movements, and shapes of the fingers. Therefore, it is necessary to improve the recognition algorithms and data glove devices (e.g., to detect hand movement periods and construct more robust glove devices).

Based on the aforementioned results, we determined that



(a)

(b)

(c)

(d)

Figure 9. Example input data (only five dimensions):
(a) predict “te” as “te” correctly, (b) predict “te” as “ke” incorrectly,
(c) predict “ke” as “te” incorrectly, (d) predict “ke” as “ke” correctly.

the recognition errors largely occurred based on variance in the flexion and direction of the fingers. We also confirmed that finger expressions varied based on individual differences, which could be attributed to different home and social environments (making recognition more difficult).

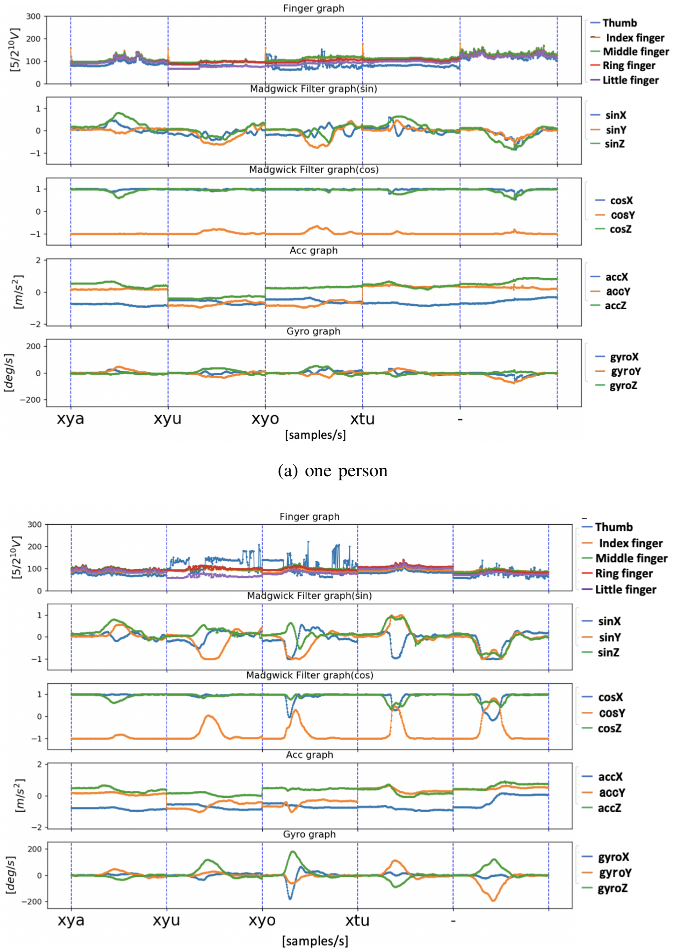
However, JF is widely used for displaying proper names and technical terms. Therefore, the recognition of JF is essential for construction a JSL recognition system.

IV. CONTINUOUS FINGERSPELLING RECOGNITION EXPERIMENT

This section describes the selection of words for data collection and for construction of a new neural network for continuous fingerspelling recognition experiments based on the system constructed in Section III. First, we describe the word selection.

A. Word selection

In a previous study [30], we proposed a method for recognizing fingerspelling words using linguistic information based on a word dictionary. We separated the recognition of actions from the recognition of hand shapes: thus, fingerspelling could be recognized even despite action recognition errors. In this experiment, we proposed 18 patterns because the number of JF patterns is more significant than those of other countries, particularly in dynamic fingerspelling, as described in Section I. Furthermore, finger and hand movements vary from person to person. Each pattern is illustrated and explained. In a previous study [30], the number of words selected was 64; thus, we selected words corresponding to that number and suitable for the 18 proposed patterns (see Figure 11). The errors were characterized as follows:



(a) one person

(b) another person

Figure 10. Example of acquiring data.

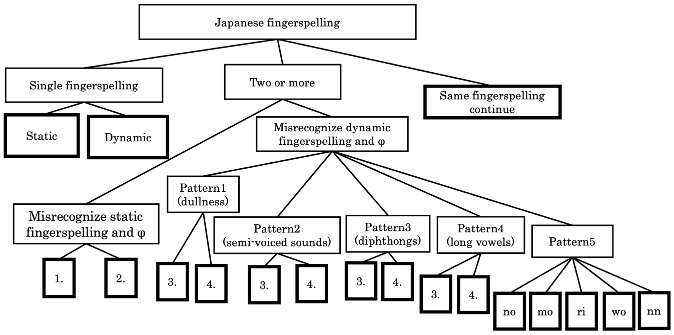


Figure 11. 64 word patterns.

1 denotes the misrecognition of static fingerspelling as the transition movements between fingerspellings;

2 denotes the misrecognition of transition movements as

static fingerspelling;

3 denotes the misrecognition of dynamic (non-static) fingerspelling as transition movements;

4 denotes the misrecognition of transition movements as dynamic fingerspelling. The tasks and groups are explained in more detail below.

Task 1: Single fingerspelling

1-1 Static fingerspelling

This comprises fingerspelling other than 1-2 dynamic fingerspelling. It is characterized by absence of hand movements.

1-2 Dynamic (non-static) fingerspelling

There are four types of dynamic fingerspellings: dullness, semi-voiced sounds, diphthongs, and long vowels. Dynamic fingerspelling is characterized by hand movements.

Task 2: Two or more fingerspellings

2-1 Misrecognizing static fingerspelling and transition movements.

2-1-1 Misrecognizing transition movements as static fingerspelling

For example, "[ta]" may be misrecognized as "[ta][ta]." The user may misrecognize "[ta]" two or more times in succession.

2-1-2 Misrecognizing static fingerspelling as transition movements

For example, "[ta][ta]" may be misrecognized as "[ta]."

2-2 Misrecognizing dynamic fingerspelling and transition movements

2-2-1 Pattern1 dullness

2-2-1-1 Misrecognizing transition movements as dynamic fingerspelling

For example, "[ta][da]" may be misrecognized as "[da]."

2-2-1-2 Misrecognizing dynamic fingerspelling as transition movements

For example, "[da]" is misrecognized as "[ta][da]."

2-2-2 Pattern2 semi-voiced sounds

2-2-2-1 Misrecognizing transition movements as dynamic fingerspelling

For example, "[pa][pa]" may be misrecognized as "[pa]."

2-2-2-2 Misrecognizing dynamic fingerspelling as transition movements

For example, "[pa]" may be misrecognized as "[pa][pa]."

2-2-3 Pattern3 diphthongs

2-2-3-1 Misrecognizing transition movements as dynamic fingerspelling

For example, "[tsu][tsu]" may be misrecognized as "[tsu][xtsu][tsu]."

2-2-3-2 Misrecognizing dynamic fingerspelling as transition movements

For example,"[tsu][xtsu][tsu]" may be misrecognized as "[tsu][tsu]."

2-2-4 Pattern4 -(long vowels)

2-2-4-1 Misrecognizing transition movements as dynamic fingerspelling

For example, "[hi][-(long vowel)]" may be misrecognized as "[- (long vowel)]."

2-2-4-2 Misrecognizing dynamic fingerspelling as transition movements

For example, "[- (long vowel)]" may be misrecognized as "hi- (long vowel)."

2-2-5 Pattern5

2-2-5-1 Misrecognizing "[no]" as transition movements

For example, "[no]" may be misrecognized as "[hi][no]."

2-2-5-2 Misrecognizing "[mo]" as transition movements

For example, "[mo]" may be misrecognized as "[to][mo]."

2-2-5-3 Misrecognizing "[ri]" as transition movements

For example, "[ri]" may be misrecognized as "[u][ri]."

2-2-5-4 Misrecognizing "[wo]" as transition movements

For example, "[wo]" may be misrecognized as "[o][wo]."

2-2-5-5 Misrecognizing "[nn]" as transition movements

For example, "[nn]" may be misrecognized as "[hi][nn]."

Task 3: Identical fingerspelling problems

For example, "[ta][ta]" may be misrecognized as "[ta]."

The words are selected using the 18 patterns described in Figure 11. Because finger and hand movements vary from person to person, we organized the words into 26 groups. Table IV shows the fingerspelling of each group.

Group 1: [no][u][ni][xyu][u], [u][ri][xyu][u], [ri][yu][u]

We need to identify "[u][ni]", but it may be "[u][ri]". The transitions from "[u]" to "[ni]" and from "[u]" to "[ri]" have similar hand movements, with there being significant difference in the speed of the fingers. In this experiment, we will collect data on three words, "[u][ri][xyu][u]", "[no][u][ni][xyu][u]", and "[ri][yu][u]", and investigate the differences in speed.

Group 2: [su][u][ri][xyo][u], [su][ri][yo][u]

Although the subject of the experiment can correctly express "[su][u][ri]", the possibility that he will misrecognize it as "[su][ri]" exists. The words "[u]" and "[ri]" are difficult to distinguish and have the same finger positions, and depend on whether hand movements are used or not. Therefore, we can expect that "[u]" may be recognized as transition movements, which would imply "[su][ri]." We will use the two-word data of "[su][ri][yo][u]" and "[su][u][ri][xyo][u]" in the recognition experiment.

Group 3: [ru][-][ru], [ru][ru]

Many people expressing "[ru][-][ru]" do not express long vowels. Therefore, a possibility exists of misrecognizing "[ru][-][ru]" as "[ru][ru]". In this experiment, we conduct a recognition experiment using two sets of data: "[ru][-][ru]" and "[ru][ru]". We then investigate the need to distinguish the differences between "[ru][-][ru]" and "[ru][-][ru]" to realize a fingerspelling recognition system.

Group 4: [su][su][gi], [su][zu][ki]

The word "[su][su][gi]" is expressed using a downward extension of the thumb, index finger, and middle finger. In addition to the use of "[su]" twice, we also investigate whether the participants could discriminate between the two words "[su][su][gi]" and "[su][zu][ki]" by adding a murmur. This task enables us to consider the algorithms necessary for obtaining sufficient information from moving fingers.

TABLE IV. 64 WORDS INTO 26 GROUPS

Item	Group	Word	1-1	1-2	2-1-1	2-1-2	2-2-1-1	2-2-1-2	2-2-2-1	2-2-2-2	2-2-3-1	2-2-3-2	2-2-4-1	2-2-4-2	2-2-5-1	2-2-5-2	2-2-5-3	2-2-5-4	2-2-5-5	3
1	1	[no][u][ni][xyu][u]	*			*														
2		[u][ri][xyu][u]	*	*													*			
3		[ri][yu][u]	*	*													*			
4	2	[su][u][ri][xyo][u]	*	*		*														
5		[su][ri][yo][u]	*	*	*															
6	3	[ru][-][ru]	*	*										*						
7		[ru][ru]	*																	*
8	4	[su][su][gi]	*	*																*
9		[su][zu][ki]	*	*				*												
10	5	[hu][re][-][mu]	*	*					*					*						
11		[pu][re][mu]	*	*						*			*							
12	6	[tsu][tsu][mi]	*			*	*													*
13		[tsu][du][mi]	*	*			*	*												
14		[chi][di][mi]	*	*																
15		[tsu][mi]	*		*															
16	7	[po][tsu][ri]	*	*																
17		[po][xtsu][tsu][ri]	*	*								*								
18	8	[me][xtsu][ki]	*	*								*								
19		[me][tsu][ki]	*								*									
20	9	[hi][nn][to]	*	*									*						*	
21		[hi][-][to]	*	*										*						
22		[pi][-][to]	*	*						*										
23		[bi][-][to]	*	*						*										
24		[no][-][to]	*	*										*	*					
25	10	[ro][-][so][nn]	*	*		*														
26		[ro][-][nn]	*	*	*															
27	11	[hi][ku]	*		*															
28		[no][ku]	*	*											*					
29	12	[ka][tsu][wo]	*	*														*		
30		[ka][tsu][o]	*																	
31	13	[a][nn][za][nn]	*	*																
32		[te][nn][ke][nn]	*	*																
33		[de][nn][ge][nn]	*	*																
34	14	[ma][tsu][ya]	*																	
35		[ma][chi][ya]	*																	
36		[ma][xtsu][chi][ya]	*	*																
37	15	[ni][ba][i]	*	*				*												
38		[ni][ha][i]	*				*													
39	16	[pa][pa]		*						*										*
40		[ha][ha]	*						*											*
41	17	[mo][to][mo][to]	*	*												*				
42		[to][mo][do][mo]	*	*												*				
43	18	[ta][ta]	*			*	*													*
44		[ta][da]	*	*			*	*												
45		[da][da]		*				*												*
46	19	[ne][sa][se][ru]	*																	
47		[ne][za][sa][se][ru]	*	*				*												
48	20	[he][ya]	*						*											
49		[pe][ya]	*	*						*										
50	21	[ko][so][gu]	*	*		*														
51		[go][zo][ku]	*	*				*												
52		[go][hu][ku]	*	*				*												
53		[ko][bu][ku]	*	*				*												
54	22	[wa][ro][shi]	*																	
55		[wa][nu][shi]	*																	
56	23	[he][ya][gi]	*	*			*													
57		[be][ki]	*	*	*															
58	24	[ho][e][ki]	*				*													
59		[bo][e][ki]	*	*				*												
60	25	[shi][se][i]	*																	
61		[ji][ra][i]	*	*																
62		[shi][ze][i]	*	*																
63	26	[ka][chi][na][no][ri]	*	*																
64		[ga][i][su][u]	*	*																
		Count	62	46	5	6	7	10	3	5	1	2	2	4	2	2	2	1	1	7

Group 5: [hu][re][-][mu], [pu][re][mu]

This task investigates if the system can distinguish between "[hu][re]" and "[mu]." We want to identify "[hu][re][-][mu]," but it could be misidentified as "[pu][re][mu]" owing to fast hand movements. Therefore, we collect two types of data, "[hu][re][-][mu]" and "[pu][re][mu]," in this experiment and conducted a recognition experiment. While analyzing the hand movements, we investigate if we can effectively obtain information regarding the hand movements.

Group 6: [tsu][tsu][mi], [tsu][du][mi], [chi][di][mi], [tsu][mi]

We need to identify "[tsu][tsu][mi]", but cases exist in which "[tsu]" is used twice and the finger flexion becomes "[chi]", which is similar. Therefore, we collect data on four words to conduct a recognition experiments: "[tsu][tsu][mi]", "[tsu][du][mi]", "[chi][di][mi]", and "[tsu][mi]", and conducted a recognition experiment. We examine the hand movements for "[tsu][tsu][mi]" and "[chi][di][mi]."

Group 7: [po][tsu][ri], [po][xtsu][tsu][ri]

For group 6, the word "[tsu][xtsu][mi]" is unavailable. Hence, we prepared the words "[po][tsu][ri]" and "[po][xtsu][tsu][ri]." Thus, in Group 6, "[tsu][tsu][mi]" and "[tsu][du][mi]" are prepared to investigate their possibility of it being recognized as "[tsu][tsu][mi]" by expressing "[xtsu][tsu]" twice in succession.

Group 8: [me][xtsu][ki], [me][tsu][ki]

Unlike group 7, group 8 recognizes only one character of each of "[xtsu]" and "[tsu]." We investigate whether the group misrecognizes "[me][xtsu][ki]" as "[me][tsu][ki]." This is where dynamic fingerspelling is misrecognized as transition movements or static fingerspelling. The same is valid for hand movements where "[xtsu]" is a diphthong. We investigate whether the system can distinguish between "[me][xtsu][ki]" and "[me][tsu][ki]" by distinguishing the difference between "[xtsu]" and "[tsu]."

Group 9: [hi][nn][to], [hi][-][to], [pi][-][to], [bi][-][to], [no][-][to]

This task examines what type of information can be obtained to identify "[hi][nn][to]." The four words "[hi][-][to]," "[pi][-][to]," "[bi][-][to]," and "[no][-][to]" are chosen because they have the potential to produce the same hand movements as "[hi][-][to]." For example, in the "[hi][-]" and "[no][-]" parts, the finger flexion is the same, but the hand movements are different. After distinguishing the hand movements, we investigate whether the participants can identify the five words "[hi][nn][to]", "[hi][-][to]", "[pi][-][to]", "[bi][-][to]", and "[no][-][to]." During the expression of "[hi][-][to]", when "[hi]" ends, the fingers either turn to the right or the left. As the direction depends on the person, this is also be investigated.

Group 10: [ro][-][so][nn], [ro][-][nn]

We need to identify "[ro][-][so][nn]," but it is possible to misidentify as "[ro][-][nn]." Some difficulty seems to exist in discriminating between "[-][so][nn]" and "[-][nn]" seems to exist. The three characters "[-]", "[-][so]", and "[-][nn]" all extend the index finger in the same manner, but the hand direction and movement may differ per person. Therefore, we investigate whether the two words "[ro][-][so][nn]" and "[ro][-][nn]" can be discriminated.

Group 11: [hi][ku], [no][ku]

A possibility exists that "[hi][ku]" may be misrecognized as "[no][ku]." In this experiment, we compare the transition from "[hi][ku]" to "[ku]" and from "[no]" to "[ku]" and discover that distinguishing between "[hi][ku]" and "[no]" at the crossing portion is more challenging. We also investigate whether the discrimination between "[hi]" and "[no]" is possible.

Group 12: [ka][tsu][wo], [ka][tsu][o]

The possibility of misrecognizing "[ka][tsu][wo]" as "[ka][tsu][o]" exists. To determine if distinguishing between "[wo]" and "[o]" is possible, we investigate the misrecognition of dynamic fingerspelling as transition movements or static fingerspelling.

Group 13: [a][nn][za][nn], [te][nn][ke][nn], [de][nn][ge][nn]

We investigate "[ke]," "[te]," "[ge]," and "[te]," "[a]," and "[sa]." In this experiment, we also investigate a problem concerning dullness.

Group 14: [ma][tsu][ya], [ma][chi][ya], [ma][xtsu][chi][ya]

We investigate whether the system can identify the three lower case characters "[tsu]," "[chi]," and "[xtsu]." The hand movements of certain people may be difficult to distinguish when they transition from "[ma]" to "[tsu]," "[chi]", and "[xtsu]." In this experiment, we collect the data of the three characters for a recognition experiment and investigate whether they are affected by the person.

Group 15: [ni][ba][i], [ni][ha][i]

"[ni]" and "[ha]" have the same finger flexion but different hand directions. Similarly, "[ni][ba][i]" and "[ni][ha][i]" have the same finger flexion but different hand direction. In this experiment, we collect the data of the two words for the recognition experiment and investigate whether distinction between the direction, dullness, and transition movements is possible.

Group 16: [pa][pa], [ha][ha]

We investigate whether "[pa][pa]" can be misrecognized as "[ha][ha]". This is a task in which dynamic fingerspelling is misrecognized as transition movements. This experiment specifically examines whether dynamic fingerspelling is misrecognized as "[ha][ha]" or "[pa][pa]." The finger movement may cause transition movements, which may lead to the misrecognition of "[pa]" as "[ha][ha]." Therefore, we collect by collecting "[ha][ha]" and "[pa][pa]" data.

Group 17: [mo][to][mo][to], [to][mo][do][mo]

There is a high possibility that the discrimination between "[mo]" and "[to]" will be difficult. We investigate whether it is possible to discriminate between "[to]" and "[mo]" using a series of fingerspelling tasks where it is possible to discriminate only one character. It is essential to focus on the speed, as it may be affected by the person.

Group 18: [ta][ta], [ta][da], [da][da]

We investigate three forms of "[ta][da]" two consecutive forms, i.e., the alternating forms of "[ta][da]" and "[da]" and two consecutive forms of "[da]." We investigate whether the the system can distinguish between "[ta]" and "[da]" as "[ta]" and "[da]" must each be expressed once. The possibilities of misrecognizing "[ta][ta]" as "[ta][da]" or "[da][da]" and vice versa are investigated.

Group 19: [ne][sa][se][ru], [ne][za][sa][se][ru]

The first word contains "[sa]" and the second does "[za]." The task is to distinguish the difference between transition movements and dullness and through this experiment, we investigate whether the system can distinguish between "[sa]" and "[za][sa]."

Group 20: [he][ya], [pe][ya]

The possibility of misrecognizing "[he][ya]" as "[pe][ya]" exists. The necessary information on the hand movements is efficiently obtained by comparing the hand movements of transitions from "[he]" to "[ya]" and from "[pe]" to "[ya]." After collecting the data on "[he][ya]" and "[pe][ya]," we conduct a recognition experiment to determine whether human factors affect the results.

Group 21: [ko][so][gu], [go][zo][ku], [go][hu][ku], [ko][bu][ku]

"[go][zo][ku]" can be misidentified as "[ko][bu][ku]", "[ko][so][gu]", or "[go][hu][ku]." We believe that the inclusion of dynamic fingerspelling in the transition movements complicate the identification of the word as "[so]," "[bu]," or "[zo]." In this experiment, we investigate if we can discriminate between "[go][zo][ku]," "[ko][bu][ku]," "[ko][so][gu]," and "[go][hu][ku]."

Group 22: [wa][ro][shi], [wa][nu][shi]

"[wa][nu][shi]" may be misrecognized as "[wa][ro][shi]." We investigate whether the difference between "[ro]" and "[nu]" can be correctly identified. In particular, we investigate the degree of discrimination between the transitions from "[wa]" to "[ro]" and from "[wa]" to "[nu]."

Group 23: [he][ya][gi], [be][ki]

"[be][ki]" can be misrecognized as "[he][ya][ki]." In this experiment, we investigate whether "[be][ki]" is misrecognized as "[he][ya][ki]" by transitioning from "[be]" to "[ki]." We also consider the effects of different signers.

TABLE V. JAPANESE FINGERSPELLING COUNT

fingerspelling	a	i	u	e	o
count	1	6	6	2	1
fingerspelling	ka	ki	ku	ke	ko
count	3	6	5	1	2
fingerspelling	sa	shi	su	se	so
count	2	4	5	3	2
fingerspelling	ta	chi	tsu	te	to
count	2	4	8	1	7
fingerspelling	na	ni	nu	ne	no
count	1	3	1	2	4
fingerspelling	ha	hi	hu	he	ho
count	2	3	2	2	1
fingerspelling	ma	mi	mu	me	mo
count	3	4	2	2	2
fingerspelling	ya	yu	yo		
count	5	1	1		
fingerspelling	ra	ri	ru	re	ro
count	1	7	4	2	3
fingerspelling	wa	wo	nn		
count	2	1	6		
fingerspelling	ga	gi	gu	ge	go
count	1	2	1	1	2
fingerspelling	za	ji	zu	ze	zo
count	2	1	1	1	1
fingerspelling	da	di	du	de	do
count	2	1	1	1	1
fingerspelling	ba	bi	bu	be	bo
count	1	1	1	1	1
fingerspelling	pa	pi	pu	pe	po
count	1	1	1	1	2
fingerspelling	xya	xyu	xyo	xtsu	-[long vowels]
count	1	1	1	1	8

Group 24: [ho][e][ki], [bo][e][ki]

When transitioning from "[ho]" to "[e]," the direction of the hand changes depending on the person. When "[ho]" ends, the hand turns to the right to express it. At this time, it may become "[bo]." In this experiment, we analyze data and video recordings.

Group 25: [shi][se][i], [ji][ra][i], [shi][ze][i]

We investigate whether the three characters "[se]," "[ra]," and "[ze]" can be identified. We also investigate transition movements and dullness.

Group 26: [ka][chi][na][no][ri], [ga][i][su][u]

We investigate whether the system can correctly identify the differences between "[ka]" and "[ga]," "[chi]" and "[i]," "[na]" and "[su]," and "[ri]" and "[u]."

Table V shows the number of fingerspellings used for the 64 selected words.

B. Data collection

In the continuous fingerspelling recognition experiment, we collected data using a video recording of the hand and a sensor glove to record the bending and movements of the fingers, as shown in Figure 12. Consequently, we collected data from 33 people (nine people aged 20, 13 aged 21, eight aged 22, two aged 23, and one aged 24). As described above, there were a total of 64 words (see Table IV). There were five repetitions for each word. Eleven dimensions (five for the hand, three for the acceleration, and three for the gyro data) were used for each word. The number of samples was 120 sps × 8 s = 960 samples. The time to acquire a word was 8 s. Particularly, the time that the hand was placed on the desk before and after the word was expressed was 3 s and the time for expressing the word was 5 s, for a total of 8 s. For the word expression time in this experiment, we defined the maximum number of characters in an acquired word to be five, with each character to be expressed in approximately 1 s. A camera (*1 in Figure 12) was installed to record finger movements. During the data collection experiment, the collaborator (*2 in Figure 12) wore the sensor glove and expressed words displayed on an iPad. In addition, another camera (*3 in Figure 12) was installed to record the experiment. These experiments were conducted with approval from the Tsukuba University of Technology Research Ethics Committee (approval number: 2020-12)

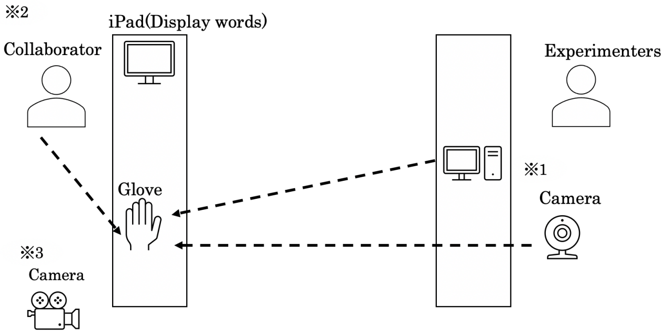


Figure 12. Data acquisition experiment of continuous Japanese fingerspelling.

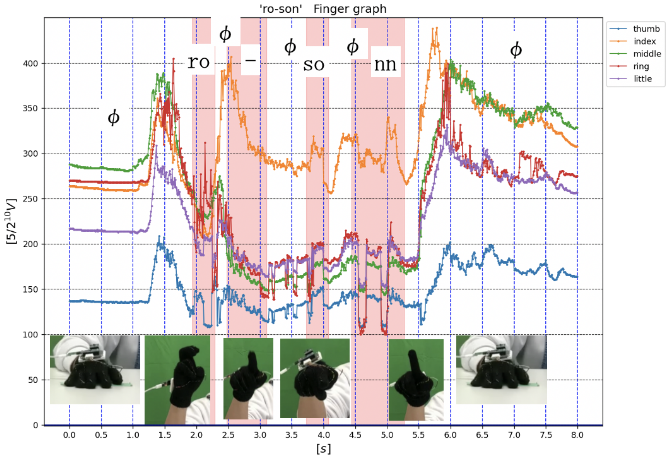


Figure 13. Five fingers of "ro-sonn."

The acceleration and gyro data collected in the experiment were used to calculate the angle using the Madgwick filter. Next, we labeled the data using ELAN software, dividing the time for each one character. For the two instances in which the hand left and was placed on the desk, and for the transition movements between characters, the blank symbol "ϕ" was inserted. To visualize the flow in a graph, "[ro][-][so][nn]" is shown as an example in Figures 13–17.

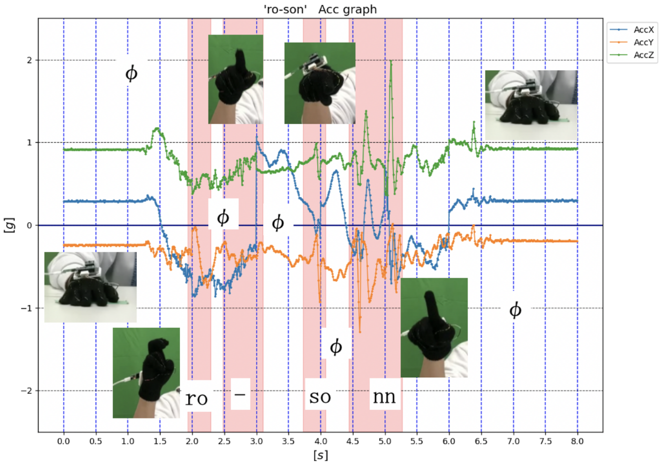


Figure 14. Accelerations of “ro-sonn.”

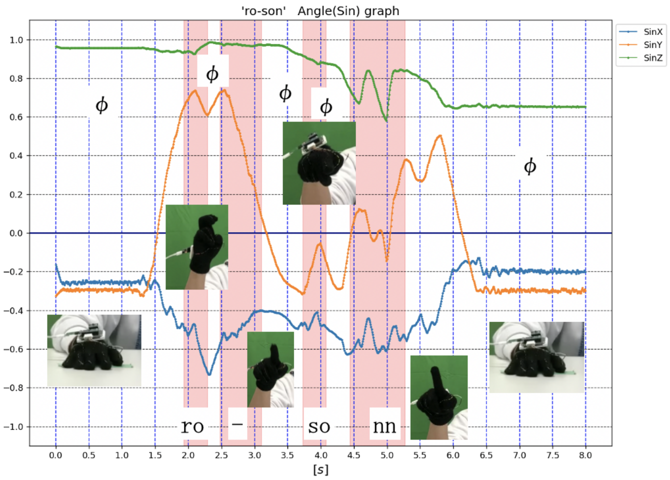


Figure 16. Sine of “ro-sonn.”

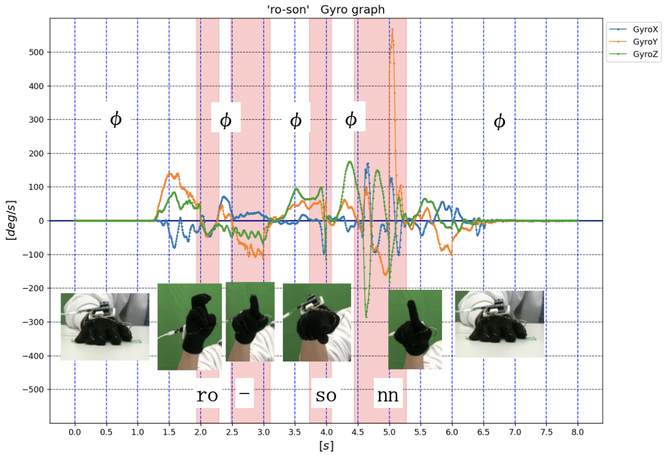


Figure 15. Gyros data of “ro-sonn.”

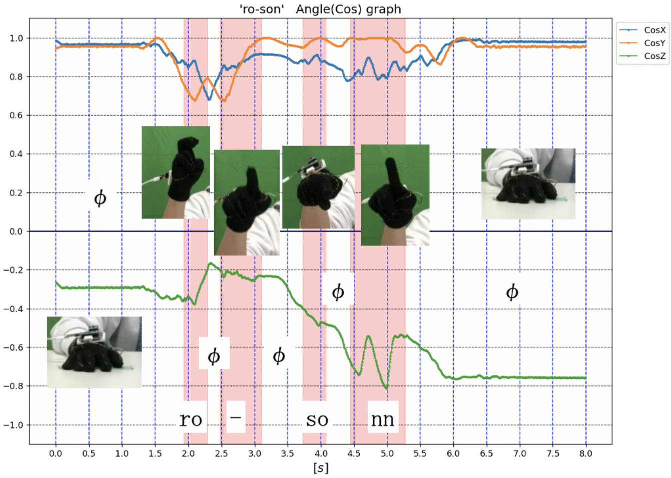


Figure 17. Cosine of “ro-sonn.”

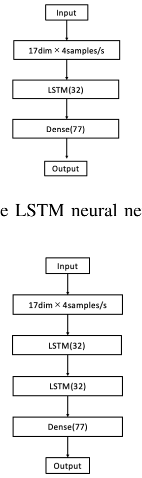
C. *Construction of neural network using long short-term memory (LSTM)*

Seven neural networks were constructed for the continuous fingerspelling recognition experiment. The LSTM [31] structure can use the short-term memory inside the network for a long time. LSTMs are often used to identify natural and speech processing language; they generally achieve high recognition rates. This experiment compares two networks, one with only the LSTM, and another with both CNN and LSTM, aiming to identify fingers, accelerations, gyro movements, and angles.

Figure 18a shows the neural network with the single LSTM as the baseline. The input data consisted of five dimensions of the hand, three dimensions of acceleration, three dimensions of gyro movements, and six dimensions of the angle, for a total of 17 dimensions × 32 samples (4 sps × 8 s). Next, the number of filters for the LSTM was 32 dimensions. In general, the number of filters is usually set to 16, 32, or 64 dimensions. Therefore, in this experiment, the number of filters of the LSTM was set to 32 dimensions, corresponding to the 77 output dimensions described below: the number of JF characters was 76, and the remaining one was “ϕ.” The latter was used to represents three situations: when the hand left the desk, when the hand was placed on the desk, and when the transition movements existed between characters.

Figure 18b shows a neural network with two LSTM layers. In this approach, the results are re-trains the results after passing through the first LSTM into the second LSTM. The number of filters in the LSTM was set to 32 dimensions, corresponding to the 17 dimensions of the hand, acceleration, gyro, and angle. The input data comprised five dimensions of the hand, three of the acceleration, three of the gyro, and six of the angle for a total of 17 dimensions × 32 samples (4 sps × 8 s). Finally, the data were output using the Dense operation (77 dimensions), i.e., the 76 JF characters and “ϕ”.

Figure 19 shows a neural network with two CNN layers and one LSTM layer. First, we input the data and then branch out into five dimensions × 32 samples (4 sps × 8 s) of the hand, three of the acceleration (4 sps × 8 s), three of the gyro (4 sps × 8 s), and six of the angle (4 sps × 8 s). After passing through the CNN (32 to 64 filters), these data were



(a) One LSTM neural network

(b) Two LSTM neural network

Figure 18. Two types of LSTM neural networks.

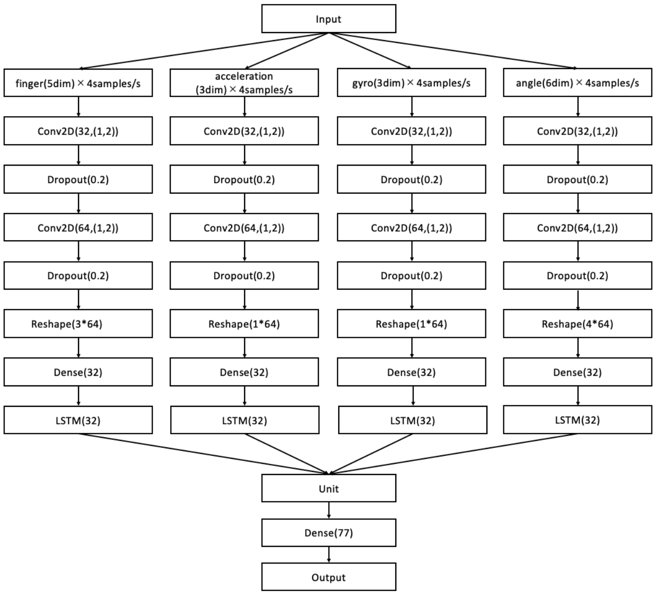


Figure 19. CNN-CNN-LSTM-unit neural network.

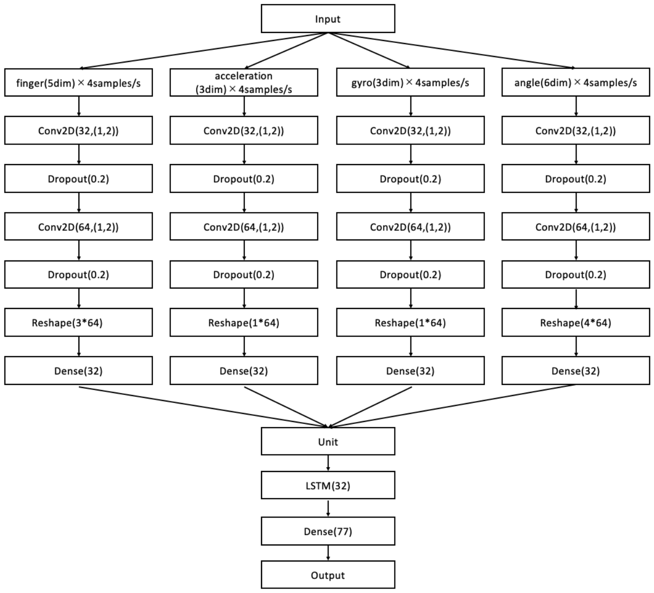


Figure 20. CNN-CNN-unit-LSTM neural network.

transformed to accommodate the Dense operation (32 nodes). Then, after passing through the LSTM (three nodes), they were combined. Finally, a Dense operation (77 nodes) corresponding to the number of characters (“ϕ” and the 76 JF characters) was applied to produce the output.

Compared to Figure 19, Figure 20 presents a different neural network in which the LSTM is added. The LSTM is interposed after combining the four datasets of the hand, acceleration, gyro, and angle, the LSTM is interposed. After inputting the data, we branched out into five dimensions × 32 samples of the hand (4 sps × 8 s), three of the acceleration (4 sps × 8 s), three of the gyro (4 sps × 8 s), and six of angle (4 sps × 8 s). These data were transformed after passing through the CNN (32 to 64 filters) to accommodate the Dense operation (32 nodes). They were then combined, and after passing through the LSTM (32 nodes), Dense operations (77 nodes) corresponding to “ϕ” and the number of characters in the JF were applied to produce the output.

Figure 21 shows the neural network with an additional LSTM for the finger, acceleration, gyro, and angle data. This re-trains the results after passing the first LSTM into the second LSTM. After inputting the data, we split the data into five dimensions × 32 samples (4 sps × 8 s of the hand, three of the acceleration (4 sps × 8 s), three of the gyro (4 sps × 8 s), and six of the angle (4 sps × 8 s). After passing through the CNN (32 to 64 filters), these data were transformed to accommodate the Dense operation (32 nodes). Then, after passing through the two LSTM layers (32 nodes), they were combined. Finally, a Dense operation (77 nodes) corresponding to the number of characters (“ϕ” and the JF characters) was applied to produce the output.

Figure 22 shows the neural network after combining the hand, acceleration, gyro, and angle data with another LSTM. First, after inputting the data, the five dimensions of hand and finger are split into 32 samples (4 sps × 8 s), three dimensions of the acceleration (4 sps × 8 s), three of the gyro (4 sps × 8 s), and six of the angle (4 sample/s × 8 s). These data were transformed after passing through the CNN (32 to 64 filters) to accommodate the Dense operation (32 nodes). They were then combined, and after passing through the two LSTM layers (32 nodes), Dense operations (77 nodes) corresponding to the number of characters in “ϕ” and the JF were applied to produce the output.

Figure 23 shows a neural network with one LSTM before and after merging. The LTSM before merging learns the results for the fingers, acceleration, gyro, and angle, after passing through the CNN. The LSTM after merging learns the results after merging the hand, acceleration, gyro, and angle datasets. First, the data ware input; then the network branched into 32

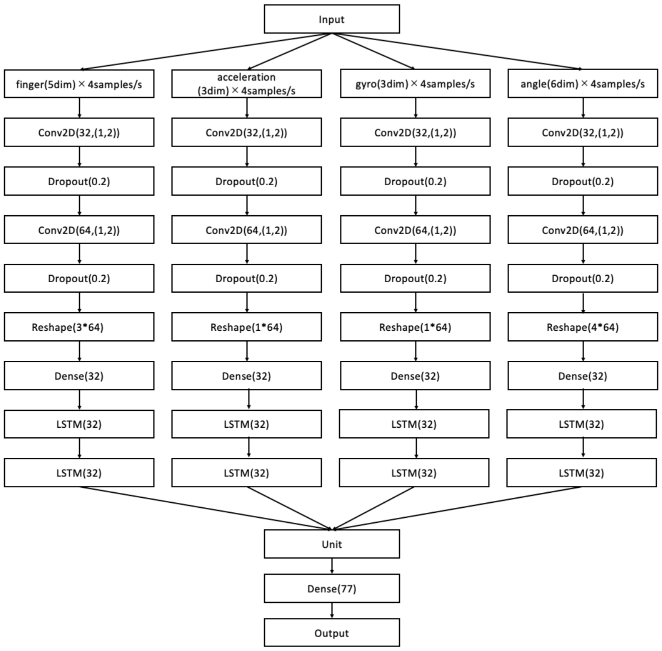


Figure 21. CNN-CNN-LSTM-LSTM-unit neural network.

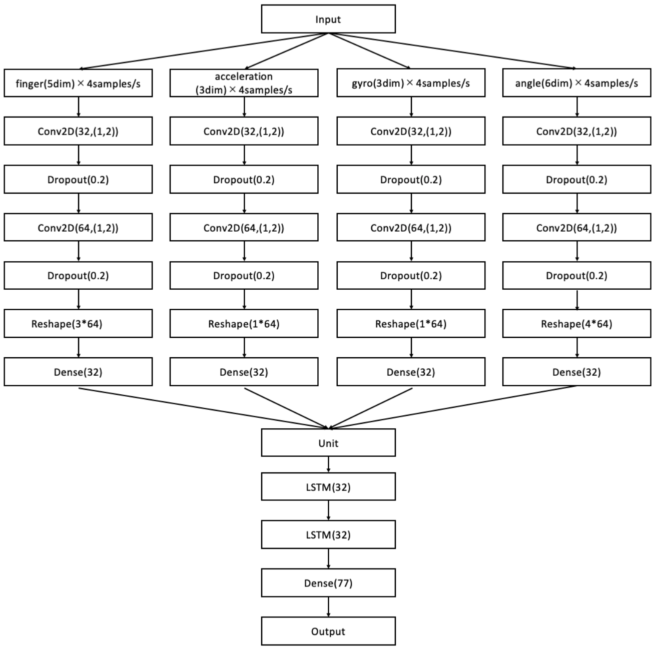


Figure 22. CNN-CNN-unit-LSTM-LSTM neural network.

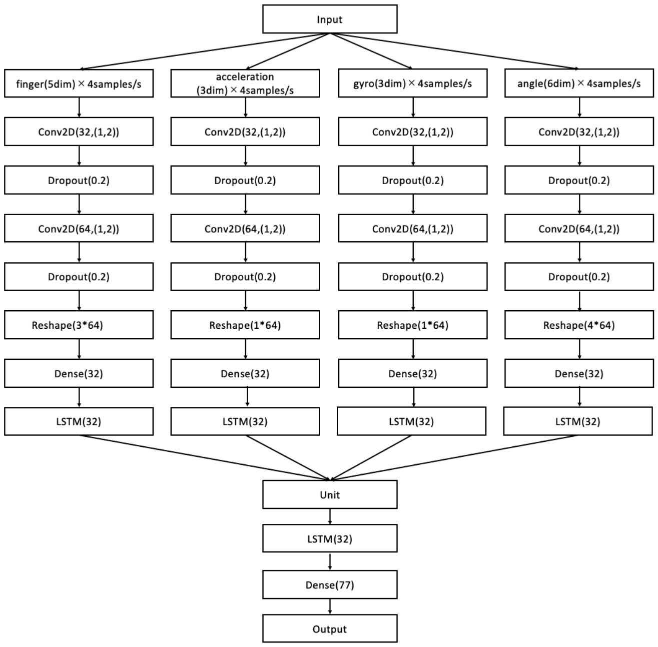


Figure 23. CNN-CNN-LSTM-unit-LSTM neural network.

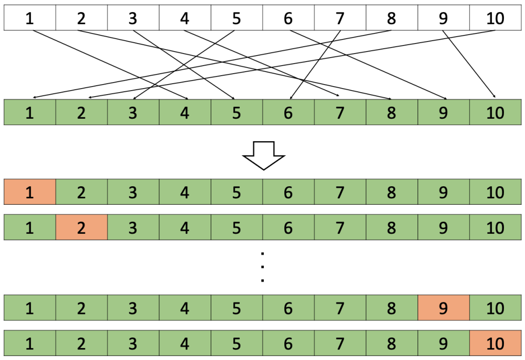


Figure 24. CNN-CNN-LSTM-unit-LSTM neural network.

samples of the five dimensions of the hand (4 sps × 8 s), three of the acceleration (4 sps × 8 s), three of the gyro (4 sps × 8 s), and six of the angle (4 sps × 8 s). These data were transformed after passing through the CNN (32 to 64 filters) to accommodate the Dense operation (32 nodes). Then, after passing through the LSTM (32 nodes), they were combined. Finally, after passing through another LSTM (32 nodes), a Dense operation (77 nodes) corresponding to the number of characters (“ϕ” and the JF characters) was applied to produce the output.

D. Evaluation experiments

We conducted evaluation experiments for each of the seven neural networks constructed in Figures 18a–23. The input data was shuffled and then divided into two parts, i.e., training and test data, using 10-fold cross-validation (see Figure 24).

E. Results and discussion

As shown in Table VI, the accuracy rates of the six neural networks, except for the one with the LSTM (one layer), are above 90%. Comparing the neural network with the two LSTM layers before merging to the neural network with the two LSTM layers after merging, the accuracy of the latter is approximately 1% higher than that of the former. In terms of good fit and recall, and the neural network with both a CNN and LSTM obtains an approximately 90% better performance

TABLE VI. SEVEN NEURAL NETWORKS EXPERIMENT RESULTS.

Neural network	Learning data (%)	Validation data (%)
one LSTM	91.4	90.1
two LSTM	92.5	90.3
branch-CNN-CNN-LSTM-unit	95.1	91.7
branch-CNN-CNN-unit-LSTM	94.0	91.8
branch-CNN-CNN-LSTM-LSTM-unit	96.5	91.3
branch-CNN-CNN-unit-LSTM-LSTM	94.7	92.1
branch-CNN-CNN-LSTM-unit-LSTM	95.2	91.6

TABLE VII. FIVE-FOLD CROSS-VALIDATION RESULTS.

k	Learning data (%)	Validation data (%)
1	95.0	92.4
2	94.7	92.0
3	94.9	92.4
4	94.5	91.8
5	94.4	91.8
Average	94.7	92.1

than the neural network with only one LSTM. That is, the neural network using both a CNN and LSTM has a higher accuracy rate. The higher accuracy may be owing to have been obtained because of the branching of the fingers, acceleration, gyro, and angle datasets and the detection of the feature points from the input data using CNNs. Therefore, we analyze the branch→Conv2D→Conv2D→conjoin→LSTM→LSTM neural network (see Figure 22) with the highest accuracy among the seven neural networks. Table VII summarizes the results of the five-fold cross-validation recognition experiments.

In this experiment, macro-averages are obtained for multi-class classification. Precision and recall are calculated using true positive (TP), false positive (FP), and false negative (FN) as shown below. The F-measure is the harmonic mean of the two values.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

Table VIII shows that the precision, recall, F-measure for macro-averages, and F-measure for micro-averages were 67.8%, 62.3%, 64.7%, and 92.1%, respectively. Table X shows the fingerspellings ranked as ordered from the smallest F-measure. The precision for "ϕ" is 95.6%, the recall is 96.6%, and the F-measure is 96.1%, i.e., relatively high value. "ϕ" may have been misrecognized as "[te]" while the fingers were motionless during data collection. The F value (35.1%) is the smallest for "[te]" among the 77 types of data. The precision is 43.1%, and the recall is 29.6%. We confirmed that "[te]" was mistaken as "[de]," "[wa]," and "ϕ" using a confusion matrix.

The F-measure of "[ho]" is 38.5%. We confirmed that "[ho]" is included in "[bo]" and "[po]" using a confusion matrix. "[ho]" is expressed with the front of the hand forward, and the space between the five fingers close together. For "[bo]", right-handed users move their fingers to the right, and left-handed users to the left after expressing "[ho]." "[po]" is expressed by fingers moving upward after expressing "[ho]." The misrecognition of "[ho]" as two characters, "[bo]" and "[po]," owes to the similarities in hand movement.

TABLE VIII. MACRO-AVERAGE AND MICRO-AVERAGE

macro			micro
‘ Precision(%)	Recall(%)	F-measure(%)	F-measure(%)
67.8	62.3	64.7	92.1

The F-measure of "[chi]" is 41.4%. The confusion matrix confirms that "[chi]" is included in "[di]," "[tsu]," "[du]," and "[xtsu]." The "[chi]" is expressed with the thumb touching the index, middle, and ring fingers and the little finger extended. For "[di]," right-handed users move their hand to the right, and left-handed users to the left, after expressing "[chi]". The differences in hand movement cause misrecognition. The misrecognition of "[tsu]," "[du]," and "[xtsu]" is caused by the ring finger being extended for "[chi]" or not.

The F-measure of "[pe]" is 45.9%. "[pe]" is included in "[he]" and "[be]," as confirmed using a confusion matrix. "[he]" is expressed with fingers pointed downward, with the thumb and little finger extended and the other three fingers flexed. "[pe]" is expressed with the fingers moving upward after expressing "[he]." To express "[be]," right-handed users move their fingers to the right, and left-handed users to the left, after expressing "[he]." "[pe]" is misrecognized as two characters, "[he]" and "[be]", because the first parts of "[pe]" and "[be]" are expressed the same as "[he]," resulting in misrecognition owing to the similarities in hand movements.

The F-measure of "[du]" is 46.2%, and the confusion matrix confirms "[du]" is included in "[tsu]" and "[di]." "[tsu]" is expressed with the thumb touching the index and middle fingers and the ring and little fingers extended. Right-handed users express "[du]" by moving their fingers to the right, and left-handed users to the left, after expressing "[tsu]." "[du]" is misrecognized as "[tsu]" owing to movement similarities.

The F-measure of "[xyo]" is 77.9%. A confusion matrix confirms that "[yo]" is included in "[yo]." For "[yo]," only the thumb is flexed, and the other four fingers are extended. Therefore, the misrecognition of "[xyo]" as "[yo]" is caused by hand movement problems.

The F-measure of "[nn]" is 79.1%. We confirmed that "[nn]" is included in "[so]" and "[-](long vowel)" using a confusion matrix. For "[nn]", the index finger is extended to express a writing image, such as "[nn]" in katakana. "[so]" is expressed with the index finger extended downward and slightly diagonal. The misrecognition of "[nn]" as "[so]" is caused by the similarities in hand movement. T he "[-]" sound is represented by the index finger extending and moving up and down. "[nn]" is misrecognized as a "[-]" because the movement required to express "[nn]" halfway resembles the movement required to express "(-)".

The F-measure of "[mu]" is 80.2%. A confusion matrix confirms that "[mu]" is included in "[ku]," using a confusion matrix. "[mu]" is expressed with the thumb and index finger extended. "[ku]" is expressed with all five fingers extended and in close contact (excepting the thumb). "[mu]" is misrec-

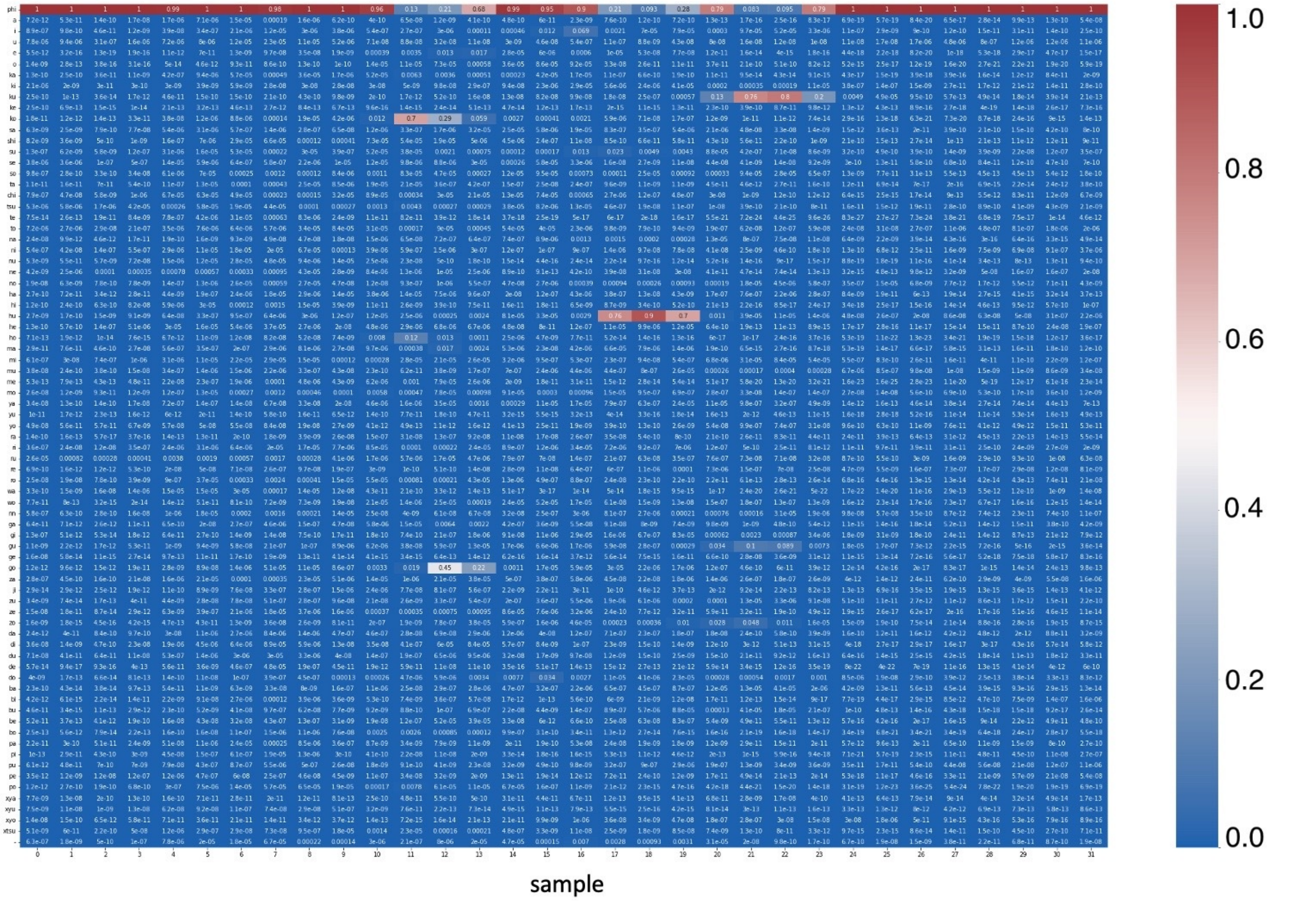


Figure 25. Recognition time series of "gohuku."

TABLE IX. "gohuku" accuracy(%)

sample	9	10	11	14	15	16	18	19	20
ko	69.0	13.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
go	3.9	77.0	83.0	0.0	0.0	0.0	0.0	0.0	0.0
hu	0.0	0.0	0.0	19.0	84.0	1.8	0.0	0.0	0.0
bu	0.0	0.0	0.0	1.0	1.0	22.0	0.0	0.0	0.0
ku	0.0	0.0	0.0	0.0	0.0	0.0	0.0	67.0	90.0
phi	19.0	7.5	14.0	79.0	12.0	76.0	95.0	32.0	9.3

ognized as "[ku]" owing to the positions of the middle, ring, and little fingers.

The F-measure of "[mi]" is 82.0%, i.e., the highest among the 76 characters other than "ϕ". The confusion matrix confirms that "[mi]" is included in "[shi]." The "[mi]" character is expressed with the index, middle, and ring fingers extended to the left for right-handed users and to the right for left-handed users. The index and middle fingers are extended to express "[shi]" with the right-handed fingers pointing left and left-handed fingers right. The misrecognition of "[mi]" as "[shi]" is caused by the position of the thumb and ring finger.

As an example of a problem in hand movement, we take the word "[go][hu][ku]." When users attempt to express "[go]," they may first express "[ko]," and we assume that it is recognized correctly. As an example, the recognition result of "[go][hu][ku]" is shown in the Figure 25. In addition, Table IX shows accuracy rate (unit: %) for each sample of "[go][hu][ku]." In this case, the user expression of the word "[go]" first expresses "[ko]" using static fingerspelling, and then performs an action. That is, the first movement of "[go]" is regarded as a static fingerspelling and become "[ko][go]." To recognize "[go]" clearly, specifying the range of time from when the finger begins moving after expressing "[ko]" to when the movement ends is necessary. In addition, it is necessary to insert a new transition movement between "[ko]" and "[go]."

V. Conclusions and Future work

In this study, to realize smooth communication between DHH and hearing people, we adapted a lightweight sensor

TABLE X. FINGERSPELLING RANK FROM THE LEAST F-MEASURE

Rank	JF	Data	Precision(%)	Recall(%)	F-measure(%)
1	te	159	43.1	29.6	35.1
2	ho	182	41.7	35.7	38.5
3	chi	676	43.9	39.2	41.4
4	pe	245	58.1	38.0	45.9
5	du	255	47.3	45.1	46.2
6	ko	360	55.5	44.7	49.5
7	di	259	53.0	47.1	49.9
8	pi	241	63.8	42.3	50.9
9	ji	256	65.1	42.2	51.2
10	so	338	55.7	47.3	51.2
11	ga	261	54.7	48.7	51.5
12	bo	252	60.6	48.8	54.1
13	ka	555	57.0	52.3	54.5
14	tsu	1801	58.7	51.5	54.8
15	ne	356	60.8	50.6	55.1
16	hu	367	56.1	55.3	55.7
17	ni	546	54.7	58.4	56.5
18	ro	509	61.9	52.7	56.9
19	ha	639	60.6	53.7	56.9
20	he	394	56.0	58.1	57.0
21	pu	242	62.1	52.9	57.1
22	hi	553	57.2	57.7	57.4
23	me	328	63.5	53.7	58.2
24	pa	437	65.9	52.6	58.5
25	xtsu	701	65.2	53.4	58.7
26	po	485	63.8	51.3	58.8
27	wo	273	63.6	56.4	59.8
28	wa	382	61.7	60.2	60.9
29	o	264	66.8	58.0	62.1
30	na	154	60.5	65.6	62.9
31	ra	189	70.6	57.1	63.2
32	su	1082	63.6	63.1	63.4
33	nu	201	70.3	57.7	63.4
34	ma	543	62.6	65.7	64.2
35	shi	863	67.2	61.4	64.2
36	de	222	70.9	60.4	65.2
37	bi	244	70.0	61.1	65.2
38	zo	248	66.0	64.9	65.4
39	go	516	67.1	64.5	65.8
40	ru	1305	65.2	67.1	66.1
41	a	170	70.4	62.9	66.5
42	ta	606	72.9	61.6	66.7
43	xya	268	69.5	64.6	66.9
44	sa	387	68.8	65.4	67.0
45	se	563	68.0	67.3	67.7
46	yu	195	70.6	65.1	67.7
47	e	365	68.3	67.4	67.9
48	ke	168	67.8	69.0	68.4
49	ba	260	68.5	68.5	68.5
50	gu	284	72.8	66.9	69.7
51	bu	260	77.5	63.5	69.8
52	yo	188	73.2	69.7	71.4
53	da	761	72.5	70.8	71.7
54	be	207	76.5	67.6	71.8
55	xyu	488	75.3	69.5	72.3
56	mo	847	75.3	69.8	72.4
57	ze	260	75.8	70.0	72.8
58	ya	1195	74.8	73.8	74.3
59	no	976	75.9	72.8	74.3
60	ki	1374	75.6	73.1	74.3
61	zu	275	76.5	72.4	74.4
62	u	1884	75.8	73.0	74.4
63	re	353	74.3	75.4	74.8
64	za	473	79.5	71.5	75.3
65	do	259	78.9	72.2	75.4
66	ge	228	80.8	71.9	76.1
67	to	1738	78.2	74.3	76.2
68	gi	550	76.2	76.7	76.4
69	i	1428	77.3	76.3	76.8
70	-(long vowel)	2115	75.4	79.0	77.1
71	ri	1890	79.3	76.3	77.8
72	xyo	259	782.2	77.6	77.9
73	ku	1263	79.4	77.7	78.5
74	nn	2878	81.7	76.6	79.1
75	mu	494	80.7	79.8	80.2
76	mi	1021	82.7	81.6	82.0
77	phi	252947	95.6	96.6	96.1

glove, developed an effective CNN model, implemented a JF recognition system, and evaluated the performance of the developed system. JF data collection experiments with 20 participants and 76 target JF characters were repeated five times. Data were acquired at 200 sps for 11 input dimensions. Angle data were transformed by applying a Madgwick filter to gyro readings and were converted into sine and cosine spaces, thereby increasing the total number of input dimensions to 17. However, the data acquired at 200 sps contained various issues, such as noisy signals. To solve this problem, we calculated the moving averages to reduce the frequency to 4 sps. Finally, a 20-fold cross-validation evaluation was conducted. The average recognition rate was approximately 70.0%, and the maximum recognition rate was approximately 75.5%. We determined that the variance in the flexion and direction of the fingers was a significant cause of misrecognition.

We then described the results of the continuous fingerspelling recognition experiment. In daily life, finger flexion, extensions, hand directions, and movements vary considerably among people learning sign language, people using ESJ, and people using JSL. Therefore, we proposed a dataset to exploit the characteristics of JF and selected 64 words. Then, we conducted a data collection experiment. For each of the 64 words, 11 dimensions (hand: five dimensions, acceleration: three dimensions, gyro: three dimensions) were input for eight s (120 sps × 8 s = 960 samples). The data and video collections were repeated five times. Then, using the acceleration and gyro data, the angles (three dimensions) were calculated using the Madgwick filter and converted to sine and cosine values. Six dimensions were added, bringing the total number of dimensions to 17, including those of the fingers (five dimensions) and accelerations and gyro (six dimensions). Next, the data was reduced by setting the average to 32 samples (4 sps × 8 s). Finally, a discrimination experiment was conducted. We compared two neural networks, one with only an LSTM and another with both CNN and LSTM. For the neural network using both CNN and LSTM, the evaluation experiment was conducted by splitting the hand, acceleration, gyro, and angle data, and passing each through the neural network. Consequently, micro F-measure of 92.1% was obtained for the neural network using the CNN and LSTM. Although we solved the calibration problem, a hand adhesion issue remained. Furthermore, distinguishing between static and dynamic fingerspelling based on hand motion became difficult.

Thus, this system had two main problems: hand-finger adhesion and distinguishing between static and dynamic fingerspellings. In the continuous fingerspelling recognition experiment, the accuracy rate of the finger characters decreased owing to the large number of instances "ϕ". To obtain a high discrimination rate in fingerspelling recognition, we must expand the data on fingerspellings and collect more data. The amount of data for "ϕ" is considerably larger than that of the 76 characters of JF; more data facilitate distinguishing between static and dynamic fingerspelling. Distinguishing between "[ko]" and "[go]" became particularly difficult; thus, we must contemplate constructing a system that considers any hand movement as a dynamic fingerspelling. The issue is also occurred in the single fingerspelling recognition experiments, e.g., distinguishing "[te]" from "[u]," "[te]" from "[tsu]," "[te]" from "[ru]," and "[te]" from "[wa]."

To further develop sign language recognition systems, three issues must be addressed. First, a large amount of speech data exists but with insufficient sign language data. To improve the accuracy rates using deep learning, collecting more sign language data is necessary. Second, preparing data on JSL is also necessary. In daily life, variations in finger flexion, hand directions, and hand movements occur among people. In addition, we must develop a learning model suitable for JSL. A multimodal approach, concerning JSL, addresses the first issue by inputting different types of information such as finger flexion, hand directions, hand movements, and facial expressions to improve recognition. JSL uses fingers, hand directions, and hand movements as well as the upper body, head, face, and mouth to express. Therefore, constructing a suitable language model for JSL is necessary. We plan to development of a sign language recognition system able to address these three issues.

VI. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number #19K11411 and the Promotional Projects for Advanced Education and Research in NTUT. We would like to thank Editage (www.editage.com) for English language editing.

References

[1] T. Tsuchiya, A. Shitara, F. Yoneyama, N. Kato, and Y. Shiraishi, “Sensor glove approach for japanese fingerspelling recognition system using convolutional neural networks,” in Proceedings of The Thirteenth International Conference on Advances in Computer-Human Interactions (ACHI 2020), 2020, pp. 152–157.

[2] “UDtalk,” 2015, URL: https://udtalk.jp/ [retrieved: December, 2022].

[3] “KoeTra,” 2015, URL: https://www.koetra.jp/en/ [retrieved: December, 2022].

[4] “speech-to-text,” 2019, URL: https://cloud.google.com/speech-to-text [retrieved: December, 2022].

[5] S. Gattupalli, A. Ghaderi, and V. Athitsos, “Evaluation of deep learning based pose estimation for sign language recognition,” in Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2016, pp. 1–7.

[6] M. A. Jalal, R. Chen, R. K. Moore, and L. Mihaylova, “American sign language posture understanding with deep neural networks,” in 2018 21st International Conference on Information Fusion (FUSION). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), July 2018, pp. 573–579.

[7] G. Luzhnica, J. Simon, E. Lex, and V. Pammer, “A sliding window approach to natural hand gesture recognition using a custom data glove,” in 2016 IEEE Symposium on 3D User Interfaces (3DUI). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), March 2016, pp. 81–90.

[8] K. Murakami and H. Taguchi, “Gesture recognition using recurrent neural networks,” in Proceedings of the SIGCHI conference on Human factors in computing systems, 1991, pp. 237–242.

[9] N. Mukai, N. Harada, and Y. Chang, “Japanese fingerspelling recognition based on classification tree and machine learning,” in 2017 Nicograph International (NicoInt). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), June 2017, pp. 19–24.

[10] H. Hosoe, S. Sako, and B. Kwolek, “Recognition of jsl finger spelling using convolutional neural networks,” 05 2017, pp. 85–88.

[11] M. E. Cabrera, J. M. Bogado, L. Fermin, R. Acuna, and D. Ralev, “Glove-based gesture recognition system,” in Adaptive Mobile Robotics. World Scientific, 2012, pp. 747–753.

[12] C. K. Mummadi, F. P. P. Leo, K. D. Verma, S. Kasireddy, P. M. Scholl, and K. Van Laerhoven, “Real-time embedded recognition of sign language alphabet fingerspelling in an imu-based glove,” in Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction, ser. iWOAR ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3134230.3134236

[13] R. Takada, J. Kadomoto, and B. Shizuki, “A sensing technique for data glove using conductive fiber,” in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI EA ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–4. [Online]. Available: https://doi.org/10.1145/3290607.3313260

[14] M. Kondo, N. Kato, K. Fukui, and A. Okazaki, “Development and evaluation of an interactive training system for both static and dynamic fingerspelling using depth image,” IEICE technical report, vol. 114, no. 512, 2015, pp. 23–28, (in Japanese).

[15] “5DT Data Glove 5 Ultra,” 2019, URL: https://5dt.com/ [retrieved: December, 2022].

[16] N. M. Kakoty and M. D. Sharma, “Recognition of sign language alphabets and numbers based on hand kinematics using a data glove,” Procedia Computer Science, vol. 133, 2018, pp. 55–62.

[17] T.-W. Chong and B.-J. Kim, “American sign language recognition system using wearable sensors with deep learning approach,” The Journal of the Korea Institute of Electronic Communication Sciences, vol. 15, no. 2, 2020, pp. 291–298.

[18] X. Yu, S. Liu, W. Fang, and Y. Zhang, “Research and discovery of smart dumb gloves,” in Journal of Physics: Conference Series, vol. 1865, no. 4. IOP Publishing, 2021, p. 042054.

[19] O. Glauser, S. Wu, D. Panozzo, O. Hilliges, and O. Sorkine-Hornung, “Interactive hand pose estimation using a stretch-sensing soft glove,” ACM Transactions on Graphics (TOG), vol. 38, no. 4, 2019, pp. 1–15.

[20] H. Zhou, W. Zhou, Y. Zhou, and H. Li, “Spatial-temporal multi-cue network for continuous sign language recognition,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 13 009–13 016.

[21] “PWTH-PHOENIX-Weather 2014-T),” 2019, URL: https://www-i6.informatik.rwth-aachen.de/ koller/RWTH-PHOENIX-2014-T/ [retrieved: December, 2022].

[22] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, “Fingerspelling recognition in the wild with iterative visual attention,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[23] “Chicago Fingerspelling in the Wild Data Sets (ChicagoFSWild, ChicagoFSWild+),” 2019, URL: https://home.ttic.edu/ klivescu/ChicagoFSWild [retrieved: December, 2022].

[24] D. Bragg, O. Koller, N. Caselli, and W. Thies, “Exploring collection of sign language datasets: Privacy, participation, and model performance,” in The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, ser. ASSETS ’20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3373625.3417024

[25] “TensorFlow,” 2019, URL: https://www.tensorflow.org [retrieved: December, 2022].

[26] “scikit-learn,” 2019, URL: https://scikit-learn.org/stable/index.html [retrieved: December, 2022].

[27] G. Hinton, N. Srivastava, and K. Swersky, “Lecture 6e-rmsprop: Divide the gradient by a running average of its recent magnitude. cousera neural networks machine learning, 2012.”

[28] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, “Estimation of imu and marg orientation using a gradient descent algorithm,” in 2011 IEEE International Conference on Rehabilitation Robotics. New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), June 2011, pp. 1–7.

[29] E. Bisong, “Google colaboratory,” in Building Machine Learning and Deep Learning Models on Google Cloud Platform. Springer, 2019, pp. 59–64.

[30] K. Kazama, Y. Horiuchi, S. Masayoshi, and S. Kuroiwa, “Continuous finger spelling recognition using kinect based on linguistic information,” IEICE technical report, vol. 117, no. 502, 2018, pp. 83–88, (in Japanese).

[31] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, 1997, pp. 1735–1780.

A Method for Analyzing Improper Driving Using Passenger's Danger Perceptions and its Evaluation

Short paper

Kensho Nishizawa
Graduate School of Engineering and Science
Shibaura Institute of Technology
Tokyo, Japan
Email: ma21111@shibaura-it.ac.jp

Tsuyoshi Nakajima
Department of Computer Science and Engineering
Shibaura Institute of Technology
Tokyo, Japan
Email: tsnaka@shibaura-it.ac.jp

Abstract—**One of the main causes of traffic accidents is "improper driving", such as driver's carelessness and operation mistakes. To prevent traffic accidents, it is necessary to detect the occurrence of improper driving, point it out to the driver, and advise improvement. However, the driving behavior analysis, which analyzes the driving itself from several sensor data, cannot accurately and comprehensively detect improper driving because it does not consider the traffic situation related to the road, the other vehicles and so on. This paper proposes a method that combines the driving behavior analysis and the danger perception by passengers. This method allows for comprehensive and correct detection of improper driving because passengers can objectively see both the driving and traffic situation. Experiments to apply this method showed its effectiveness. Furthermore, we conducted the other experiment to compare the passenger danger perception with driver's one. The results showed that drivers and passengers find out different types of dangerous driving, and that passengers perceive the driving more objectively. The results newly show that the use of objective analysis using the passenger's heart rate is effective in providing improvement to the driver.**

Keywords-Drive Analysis; Heart Rate; Passenger.

I. Introduction

Traffic accidents are one of the many serious problems in the modern society. One of the major causes of traffic accidents is improper driving by drivers, including driver errors and carelessness [2].

Support systems that point out the occurrence of improper driving and provide advice for improvement would be effective in preventing drivers from causing accidents. Driving behavior analysis is a method that uses sensors to measure the behavior of a car and detects abnormal driving patterns as improper driving, such as rapid acceleration and meandering. However, the driving behavior analysis is not highly accurate because only a few driving patterns can be analyzed without considering traffic situation, in which the driving takes places. Therefore, it is often the case where the driving behavior analysis wrongly detects or even misses improper driving. This problem can be a major drawback for such support systems to give appropriate advice to the drivers.

To solve this problem, we proposed a method that, in addition to the driving behavior analysis, utilizes passenger's danger perception of the traffic situation [1]. This paper fleshes out the contents to provide detailed discussion and evaluation of the method. The proposed method analyzes the variation of the passenger's heart rate to detect the passenger's danger perception, which is used to compare to the abnormal driving patterns detected by the driving behavior analysis. If they are matched, the detected patterns are considered improper driving. In addition, there are cases where driving is deemed improper purely based on traffic situation, which cannot be detected by driving behavior analysis. The proposed method extracts such cases by analyzing the passenger's danger perceptions that do not match the occurrence of abnormal driving patterns. This allows for more comprehensive and accurate detection of improper driving.

To implement the proposed method, we devised a method to detect danger perception based on abnormal heart rate of the target. The criteria for determining heart rate abnormality are set through experiments.

An experiment was conducted to show that our method works effectively. The results show that the passenger's danger perceptions are useful to improve the accuracy of detecting the improper driving, and passenger-perception-only data, which are the passenger's danger perceptions without matching the abnormal driving patterns, include many cases of improper driving. Some, however, were caused by things unrelated to improper driving. These must be removed to determine passenger heart rate abnormalities as dangerous driving.

A comparative experiment was also conducted to compare the passenger danger perception with driver's one. The results showed that drivers and passengers find out different types of dangerous driving, and that passengers perceive driving more objectively. Moreover, it was found that drivers and passengers often perceive the different types of dangers, and so considering both may allow for a more comprehensive and accurate detection of improper driving.

In this paper, Section II describes existing methods for detecting improper driving and their problems. Section III proposes our method, and Section IV describes an experiment and its result to show the effectiveness of the method, and Section V discusses it. Section VI describes a comparative experiment with the driver's danger perception and its results, and Section VII discusses it. Section VIII concludes this paper and discusses future prospects.

II. Existing methods for detecting improper driving

A. *Driving behavior analysis*

Driving behavior analysis is a method that detects dangerous driving patterns by analyzing sensor data on the motion of a vehicle while driving [3][4]. The sensor data includes speed, acceleration, and angular velocity. These data are used to detect dangerous driving patterns, such as sudden braking, sudden steering, sudden acceleration, and unsteady handing. However, this method has a problem that it cannot comprehensively and accurately detect improper driving due to the following reasons:

- Not all abnormal driving patterns are covered.
- Traffic situations are never considered, and so it possibly detects incorrect improper driving.

Concerning the latter, this is because the criteria for determining whether a certain driving is improper or not vary depending on the traffic situation, in which it is taken place. That is, such criteria depend on the combination of driving pattern and the traffic situation.

B. *Driver's danger perception*

Another approach to determining improper driving is to detect the danger perception while driving from the driver's own physiological data. There exist several methods to adopt this approach, such as detecting the danger perceptions by using heart rate variability [5], monitoring some visual behaviors that characterize a driver's level of vigilance [6], and estimating the driver's emotions [7]. These methods can detect the driver's danger perception when capturing external dangers, such as unexpected events or unsafe situations.

Since the driver's perception of danger comprehensively captures all the driving situation including the driving behavior and the traffic situation, it is expected to provide more comprehensive coverage than the driving behavior analysis, such as pedestrians' sudden crossing or the other vehicles' improper driving.

On the other hand, this method has the following drawbacks:

- There are individual differences in the human heart rate and the way it changes, and these individual differences may affect the accuracy of hazard detection.
- Changes in heart rate can also occur outside of danger perception, causing false detection of danger perception.
- Because drivers perceive situations subjectively, they tend not to consider their driving to be problematic. Therefore, a driver's danger perception is not suitable for use in objective determination of improper driving.

III. Proposed method

A. *Approaches*

To solve the problems of the existing methods shown in the previous section, we adopt an approach to use the passenger's danger perception. Since the passenger's danger perception is considered more objective than the driver's, the combination with driving behavior analysis would be effective.

B. *How to determine improper driving*

The proposed method for detecting improper driving using the passenger's perception of danger is shown in Figure 1.

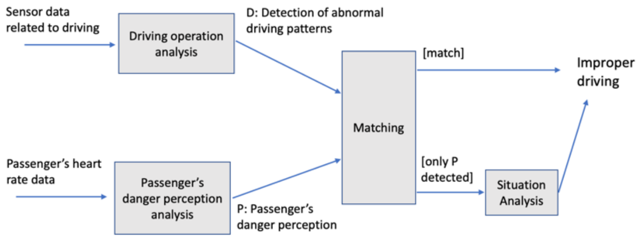

Sensor data related to driving
Driving operation analysis
D: Detection of abnormal driving patterns
[match]
Improper driving
Matching
Passenger's heart rate data
Passenger's danger perception analysis
[only P detected]
Situation Analysis
P: Passenger's danger perception

Figure 1. Proposed method for detecting improper driving using passenger's perception of danger

The method first does the driving behavior analysis using sensor data, such as Global Positioning System (GPS), speed, and angular velocity to detect abnormal driving patterns. Then, it analyzes the variation of the passenger's heart rate to detect the passenger's danger perception. After that, the detected danger perceptions are compared to the abnormal driving patterns. If they are matched, the detected patterns are determined to be improper driving. In addition, there are cases where driving is deemed improper purely based on traffic situation, which cannot be detected by driving behavior analysis. To extract such cases, the method picks up the passenger's danger perceptions that do not match the abnormal driving patterns. This allows for comprehensive and accurate detection of improper driving.

C. *Passennger's danger perception analysis*

The proposed method uses the danger perception analysis, analyzing the variation of the heart rate to find out occurrence of its abnormalities. Through an experiment, we defined the following two types of abnormal heart rate patterns.

- Rapid increase in heart rate: As shown in Figure 2, when the amount of change in heart rate increases above the specified threshold within a certain period of time, it is determined that there is a rapid increase in heart rate. The heart rate is measured every second, and the threshold of the amount of change is 10 beats.

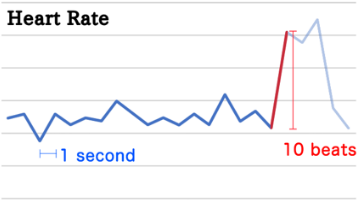


Figure 2. Heart rate variability during rapid heart rate increase

- High heart rate state: As shown in Figure 3, when the heart rate value stays high for a certain period, it is judged to be in the high heart rate state. The heart rate is judge as high when it is over the threshold of the average heart rate plus 10.

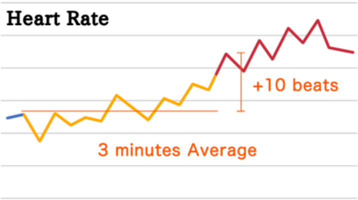


Figure 3. Heart rate variability during high heart rate state

The average heart rate is determined based on the data in the last three minutes. This compensates individual differences in the passenger's heart rate.

IV. EXPERIMENTAL EVALUATION 1

A. Purpose of the experiment

The following two hypotheses, which support the theoretical basis of the proposed method, are verified by actual driving experiments to work as expected.

- Hypothesis 1-1: False positives for improper driving will be reduced by looking at the match between abnormal driving patterns based on the driving behavior analysis and danger perception based on the passenger's danger perception analysis.
- Hypothesis 1-2: There exist some danger perceptions that do not match abnormal driving patterns, which include most of all the improper driving that strongly depends on the traffic conditions, which cannot be detected by the driving behavior analysis.

B. Methods for the experiment

In the driving experiment, the speed and angular velocity of the vehicle and the heart rate of the passenger are measured every second while driving. Each driving experiment consists of one-hour driving. After the driving, the actual driving situation was reviewed using video and notes recorded during the driving. Figures 4-7 show how the experiment took place.

- Figure 4: smartphone sensor, measuring the vehicle's angle and angular velocity.
- Figure 5: Apple Watch, measuring the passenger's heart rate.
- Figure 6: video camera on the windshield, recoding the vehicle's forward image.
- Figure 7: notes the passengers take when they perceive a danger.

In this experiment, we use a prototyping system implementing the proposed method in Figure 1, except for Situation Analysis, which is performed based on the passenger's own judgement by reviewing the recorded video after the experiment.

The subjects of this experiment are 17 university students with different driving experiences, and 50 sets of experiments took place. Table 1 shows the breakdown of the driving experience of the passengers, and Table 2 shows the breakdown of the combinations of driver and passenger by driving experience.

Figure 4. Measurement of vehicle speed and angular velocity using smartphone sensors

Figure 5. Measurement of passenger's heart rate using Apple Watch

Figure 6. Recording of the vehicle's forward image using a video camera

Figure 7. Passengers taking notes when they perceive a danger.

TABLE I. BREAKDOWN OF PARTICIPANTS IN THE DRIVING EXPERIMENT

Participants' driving frequency	Number of people	Number of experiments
Drive on a daily basis	5 people	12 sets
Sometimes drive	4 people	10 sets
Don't usually drive	4 people	13 sets
No driver's license	4 people	15 sets
Total	17 people	50 sets

TABLE II. BREAKDOWN OF DRIVERS AND PASSENGERS BY DRIVING FREQUENCY

Driver's driving frequency	Passengers' driving frequency	Number of experiments
Drive on a daily basis	Drive on a daily basis	7 sets
	Sometimes drive	1 set
	Don't usually drive	8 sets
	No driver's license	4 sets
Sometimes drive	Drive on a daily basis	5 sets
	Sometimes drive	9 sets
	Don't usually drive	5 sets
	No driver's license	11 sets
Total		50 sets

C. *Results of experiments*

1) *Verification of Hypothesis 1-1*

A total of 56 abnormal driving patterns are detected by the driving behavior analysis. Of these, 9 cases match the passenger's danger perception. Figure 8 shows all the cases classified into four danger levels by the review of actual situation.

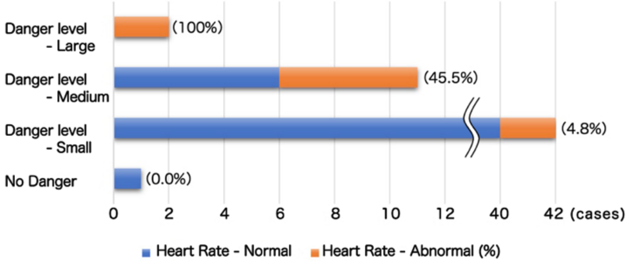


Figure 8. Relationship between abnormal driving patterns and passengers' perception of danger

Figure 8 shows that the passenger's danger perception occurs in most situations where the actual danger level is high, and in contrast, the passenger's danger perception does not occur in most situations where the actual danger level is low. There can be seen a tendency that the group of abnormal driving patterns classified in higher danger level have more percentage of ones matching with passenger's danger perception. This supports Hypothesis 1-1 that false positives for improper driving in the driving behavior analysis can be reduced by using the results of the passenger's danger perception analysis.

2) *Verification of Hypothesis 1-2*

Of the results of detecting passenger's danger perceptions, 142 cases do not match the abnormal driving patterns. Table 3 shows the results of categorizing their causes based on the review of the actual situation one by one.

TABLE III. CATEGORIZATION OF THE ANALYZED CAUSE OF PASSENGER'S HEART RATE ABNORMALITIES

Dangerous driving	Outside threats	Emotional change	Unknown cause
18 cases	43 cases	62 cases	19 cases

Details of each item used in the classification and examples of actual occurrences are as follows.

a) Dangerous driving: 18 cases

- Anxiety or fear felt about the driver's dangerous driving (e.g., accelerating instead of stopping at a traffic light change, etc.)

b) Outside threats: 43 cases

- Perceived danger due to external factors, such as interruptions by other vehicles or pedestrians jumping out (e.g., a driver suddenly getting out of a stopped truck)
- Anxiety caused by environmental factors, such as narrowness of the road and poor visibility (e.g., glare from the western sun, thick fog, etc.)

c) Emotional change: 62 cases

- Excitement or surprise during conversation
- Drowsiness and fatigue

Hypothesis 1-2 proves correct, from the fact that many improper driving cases are included in the passenger's danger perceptions that do not match abnormal driving pattern, and few improper driving cases are seen that are not detected by either analysis.

However, the results showed that there are some cases other than improper driving among the passenger's heart rate abnormalities.

V. ANALYSIS OF THE RESULTS OF EXPERIMENT 1

From the fact that the higher danger level of abnormal driving patterns matches with passenger's heart rate abnormalities well, we found that our approach to using passenger's danger perceptions to reduce the false detection of improper driving is reasonable. In addition, we also found that the passenger's danger perceptions can cover most of improper driving that the driving behavior analysis cannot detect. These results show that the proposed method can detect improper driving comprehensively and correctly.

However, the important issues to establish our method are:

- to extract only passenger's danger perceptions from all the passenger's heart rate abnormalities.
- to analyze and classify the passenger's danger perception that does not match abnormal driving patterns.

However, the above issues are difficult to settled at this moment because they require a comprehensive understanding of the factors that cause the abnormal heart rate.

VI. Experimental evaluation 2

A. *Purpose of the experiment*

Because the experiments in the previous section did not measure driver's heart rate abnormality, and therefore we could not yet prove that the passenger's heart rate abnormality is more suitable for detecting objective danger perception than the driver's one.

In this section, we set the following three hypotheses, which show the validity of the proposed method adopting the passenger's danger perception by comparison with driver's one:

- Hypothesis 2-1: Driver's heart rate abnormality matches the driver's danger perception, and a danger that the driver is unaware cannot be detected by driver danger perception.
- Hypothesis 2-2: Passengers can provide a more objective danger perception than the driver, including improper driving that the driver is unaware of.
- Hypothesis 2-3: While drivers are more careful to external dangers and can perceive more small dangers than the passengers, passengers can perceive improper driving and its situation more objectively. Therefore, the passengers can find the driver's improvement points.

B. *Methods for the experiment*

In the experiment, we used both driver's and passenger's danger perception.

The driver's and passenger's heart rate are measured using the Apple Watch's optical heart rate sensor. After the experiment, the recorded video was reviewed, and a review was conducted based on the judgment of the driver and passengers through discussion. Figures 9 and 10 show how the experiment took place.

- Figure 9: Apple Watch, measuring the driver's heart rate.
- Figure 10: Driver and passenger reviewing the video after the end of driving.

The subjects of this experiment are 6 university students with different driving experiences, and 5 sets of experiments took place. Each experiment consists of one-hour driving. Table 4 shows the breakdown of the driving experience of the passengers, and Table 5 shows the breakdown of the combinations of driver and passenger by driving experience.

Figure 9. Measurement of driver's heart rate using Apple Watch

Figure 10. Driver and passenger reviewing the video after the end of driving

TABLE IV. Breakdown of Participants in the Driving Experiment

Participants' driving frequency	Number of people
Drive on a daily basis	1 people
Sometimes drive	2 people
Don't usually drive	1 people
No driver's license	2 people
Total	6 people

TABLE V. Breakdown of Drivers and Passengers by Driving Frequency

Driver's driving frequency	Passengers' driving frequency	Number of experiments
Drive on a daily basis	Drive on a daily basis	0 set
	Sometimes drive	1 set
	Don't usually drive	0 set
	No driver's license	1 set
Sometimes drive	Drive on a daily basis	0 set
	Sometimes drive	0 set
	Don't usually drive	1 set
	No driver's license	2 sets
Total		5 sets

C. *Results of experiments*

1) *Verification of Hypothesis 2-1 & Hypothesis 2-2*

A total of 107 abnormalities were detected by the driver's and passenger's danger perception. Table 6 shows the results of classifying the causes of the abnormalities one by one by checking the actual situation, along with the presence or absence of abnormal heart rate of the driver and passenger.

TABLE VI. COMBINATION OF HEART RATE ABNORMALITIES AND CAUSE CLASSIFICATION RESULTS

Abnormal heart rate	**Cause**			
	Dangerous driving	**External factors**	**Environmental factors**	**Other**
Both	3 cases	2 cases	1 case	8 cases
Driver only	4 cases	21 cases	4 cases	39 cases
Passenger only	3 cases	4 cases	7 cases	11 cases
Total	**10 cases**	**27 cases**	**12 cases**	**58 cases**

Of the cases where heart rate abnormalities were detected, the following is a breakdown of the actual abnormal driving patterns found.

・Both parties had a heart rate abnormality: 3 cases
・Heart rate abnormality only in the driver: 4 cases
・Heart rate abnormality only in the passenger: 3 cases

Table 7 shows the breakdown of the awareness of improper driving by the driver and passenger in these 10 cases.

TABLE VII. RESULTS OF CLASSIFICATION OF HEART RATE ABNORMALITY AND AWARENESS.

Abnormal heart rate	**Awareness**		
	Both	**Driver only**	**Passenger only**
Both	2 cases	1 case	0 case
Driver only	1 case	3 cases	0 case
Passenger only	0 case	0 case	3 cases

In all cases where abnormalities were detected in the driver's heart rate, the driver himself was aware that he did improper driving. On the other hand, in cases where no abnormalities were detected in the driver's heart rate, the driver was not aware of improper driving. As shown above, the driver's heart rate abnormality has a strong correlation with the driver's awareness of improper driving.

In addition, there were 3 cases of improper driving that the driver was unaware, but the passenger felt that the driving was dangerous, and the abnormal heart rate of the passenger made it possible to detect it.

2) Verification of Hypothesis 2-3

The major causes of the driver's and passenger's heart rate abnormality that did not match abnormal driving patterns as follows.

- Driver: **external** factors: 21 cases, including other vehicles, pedestrians, or other factors may pose a danger to my vehicle, and I was worried and alarmed.
- Passenger: **environmental** factors: 7 cases, including narrow roads, poor visibility, and other factors caused dangerous driving conditions that made me feel anxious and uncomfortable.

There are 21 cases of driver's danger perception caused by external factors were detected. In many of these cases, the driver quickly predicts the possibility of danger based on the surrounding situation.

On the other hand, there are 7 cases of passenger's danger perception caused by environmental factors were detected. These cases capture passenger's anxiety and discomfort about driving in a particular traffic situation.

VII. ANALYSIS OF THE RESULTS OF EXPERIMENT 2

Firstly, 10 dangerous driving situations were detected by the driver's improper driving. Among them, the drivers were aware of their own improper driving when it was detected by the driver's heart rate abnormality, and, on the other hand, improper driving that the driver was unaware was not detected by the driver's heart rate abnormality. This supports Hypothesis 2-1 since the driver's awareness of improper driving is related to whether the driver's heart rate abnormality exists or not.

Secondly, there were 3 cases of improper driving that the driver was unaware of, but the passenger was aware of the danger. In these cases, heart rate abnormality of the passenger was detected. This supports Hypothesis 2-2 since the passenger objectively perceived the situation to recognize improper driving. The driver's danger perception detects many dangers from external factors. While driving, drivers need to pay more attention to their surroundings than their passengers. For this reason, we consider that the driver can notice existing or potential dangers caused by external factors or the possibility of such danger. On the other hand, passengers are aware of many danger perceptions from environmental factors. Thus, drivers and passengers pay attention to different things, and perceive different types of dangers.

Thirdly, we speculate that drivers are unconsciously aware of their surroundings and are alert to the possibility that an external factor may actually occur. On the other hand, passengers objectively perceive dangers in the current situation. This suggests that the passenger's danger perception is more suitable for recognizing the actual dangerous traffic situation and for providing drivers for effective advice for improvement. This supports Hypothesis 2-3.

Furthermore, since drivers and passengers perceive different types of dangers, utilizing both types of danger perceptions may allow for more comprehensive and accurate detection of improper driving.

Only five experiments were conducted, in which only a few dangers were detected. Therefore, it appears to be necessary to conduct more experiments to get more robust results in danger perception by both drivers and passengers. Concerning the subjects of our experiments, we used almost the same aged university students, but it is necessary to use a wider variety of subjects to increase the validity of the experiments.

VIII. CONCLUSION AND FUTURE OUTLOOK

In this paper, we proposed a method for detecting improper driving using driving behavior analysis, combined with the passenger's danger perception. The driving experiments showed that the proposed method can reduce the number of false positives for improper driving, and that it includes most of improper driving that the driving behavior analysis cannot detect because it strongly depends on traffic conditions. However, the passenger heart rate abnormalities include many cases not relating to improper driving, which need to be excluded.

In addition, a comparative experiment of driver's and passenger's danger perception showed that passengers can more objectively perceive improper driving that the driver is unaware of. And also, since the types of danger that drivers and passengers can perceive are different, utilizing both of the two factors may provide a more comprehensive and accurate detection of improper driving.

In the future, we will improve our method based on the findings in this paper into a comprehensive and accurate method for detecting improper driving. We also aim to implement a driving improvement support system using this method that provides effective advice to drivers.

REFERENCES

[1] K. Nishizawa and T. Nakajima, "A Method for Analyzing Improper Driving Using Passenger's Danger Perceptions", CORETA 2021 - Advances on Core Technologies and Applications, Athens, Greece, November 14 - 18, 2021

[2] Injury facts, "Motor vehicle safety issues", Available from: https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/improper-driving-and-road-rage/ , [retrieved: 12,2022]

[3] P. Wang et al., "You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis", Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2457-2466, 2018

[4] Sompo Japan Insurance Inc., "Portable smiling road navi" , Available from: https://www.sompo-japan.jp/hinsurance/smilingroad/pc/, (in Japanese) [retrieved: 12,2022]

[5] I. Kageyama, Y. Kuriyagawa, and A. Tsubouchi, "Study on Construction of Driver Model for Obstacle Avoidance Using Risk Potential", In Proceedings of the 25th International Symposium on Dynamics of Vehicles on Roads and Tracks, vol. 48, no. 2, pp. 431-437, 2017

[6] J. Wang, W. Xu, and Y. Gong, "Real-time driving danger-level prediction", Engineering Applications of Artificial Intelligence, vol. 23, no. 8, pp. 1247-1254, 2010

[7] I. Watanabe, R. Yoshida, F. Chen, and M. Sugaya, "Human Emotional State Analysis During Driving Simulation Experiment Using Bio-Emotion Estimation Method", IEEE 42nd Annual Computer Software and Applications Conference, vol. 2, pp. 589-594, 2018

Recommendations on Promoting Peer Interactions Within a Future Innovative Distance Learning Device Intended for French Orthodontic Practitioners

Contribution of a Community of Practice Analysis

Aurélie Mailloux
Orthodontics
Reims Hospital
Reims, France
e-mail: aurelie.mailloux@univ-reims.fr

Jérôme Dinet
CNRS, INRIA, Loria
University of Lorraine
Nancy, France
e-mail: jerome.dinet@univ-lorraine.fr

Abstract-**The COVID-19 crisis has changed behaviors and needs of orthodontic practitioners related to (i) cancellation of all the continuing education events, which led to the disappearance of formal and informal exchanges on the practice (ii) emergence of numerous videoconferences, but without any prior identification of practitioners' needs. The problem of interaction within a continuing education online environment is paramount: promoting interaction between peers within the system is essential to (i) reduce the feeling of loneliness (ii) promote users' commitment. Most French orthodontic practitioners were already involved in a virtual active Community of Practice (CoP) with their own way of fostering identification, cohesion, and collaboration. The purpose of this user-centered research is to identify requirements for creating an innovative comprehensive distance continuing education environment that would meet expectations and needs in terms of interactions of most CoP members, according to their experience (novices to experts). After an extensive state-of-the-art, used to better understand the changes in training and education related to orthodontic domain, we conducted (a) a detailed examination of the discursive activities within a CoP (e.g., content, interactions, rhythm, objectives) (b) four focus group (c) a survey consisting of two questionnaires (online, and face-to-face) (d) an ergonomic inspection of the *e-orthodontie.com* website. The collected data confirmed that an innovative complete distance continuing education environment could meet many CoP members' needs, such as: anonymous, scientifically validated content, extensive or limited discussion forums, clinical case sharing, videoconferences instant translation, ease of access and cost and time saving. From a theoretical point of view, this study highlighted the crucial role of the community of practice in producing requirements for creating a useful, usable, and acceptable digital education environment for orthodontic practitioners.**

Keywords-elearning; community of practice; psycho-ergonomic study; innovative device; orthodontics; continuing education.

I. INTRODUCTION

The COVID-19 crisis has changed behaviors and needs of orthodontic practitioners towards continuing education. Among others, the replacement of face-to-face congresses by videoconferences had led to the disappearance of direct formal and informal exchanges between novices and/or experts of the Community of Practice (CoP): the videoconferences current format only allows one-to-one vertical interactions between participants and speakers. However, in the field of distance continuing education, it is necessary to support a form of "e-presence" between members because one of the major dropout factors is the loneliness felt within the education device. Indeed, attrition rate is lower when the user is supported by his/her peers and interacts with them regularly [1]–[3].

The peer discussions on clinical cases could therefore help novices gain expertise (i.e., theoretical knowledge and practical skills) [4]–[6].

According to the state-of-the-art [7]–[10], several solutions are mentioned to promote interactions and commitments within an education distance device, such as distance tutoring, and e-portfolio. However, their results are heterogeneous, and their implementation complex.

This innovative continuing education environment is addressed to French orthodontic practitioners who are mostly already involved in an informal active virtual CoP, built on Facebook© in 2014. In 2022 February, this CoP was gathering almost half of the French orthodontic practitioners. The purpose of this user-centered research is to analyze requirements to promote interactions within an innovative learning system based on a dual approach. On the one hand, the CoP discursive activity was assessed quantitatively and qualitatively, and on the other hand, the CoP members' needs were identified by conducting four focus groups, a survey (which was carried out face-to-face and online), and an ergonomic inspection of the *e-orthodontie.com*, a dedicated website for orthodontic continuing education, although underused by practitioners.

The remainder of this paper is organized as follows. After a state-of-the-art (Section II), Section III describes the data gathered and methodology applied in three different studies to identify the CoP members interactions needs and attitudes according to their experience (novices to experts). This is followed by an overview of findings in Section IV, categorized by the discursive analysis, the CoP members interaction needs and the requirements. Section V summarizes the value of these findings and outlines elements of future research to be conducted on the subject.

II. State-of-the-art

We conducted an extensive state-of-the-art to identify (i) the possible benefits of designing an innovative distance learning device in the orthodontic domain (ii) the current solutions to promote interactions within distance education.

A. Interest of a "demand-pull" approach

The design industry has evolved from a "technology-push" to a "demand-pull" perspective [11]. It is now commonly known that the supply does not create its own demand because:

- from the designer's point of view, the actual uses were often disappointing
- intended and actual uses did not match
- actual uses are sometimes very heterogeneous [12]–[14].

Our user-centered design research is in line with this approach. Indeed, this study consisted in analyzing the practitioners' needs, expectations, and behaviors in terms of (i) continuing education, (ii) interactions and upstream of the design phase.

B. Contribution of an innovative device

The COVID-19 crisis has changed behaviors and needs of orthodontic practitioners. The need to shift the traditional format to remote access is now widely shared. The COVID epidemic has greatly accelerated this trend related to cancellation of all the continuing education events [15]–[17].

The state-of-the-art [11]-[18][30][31] demonstrated that many devices dedicated to the continuing education of dentists or orthodontists have been created over the past 20 years, particularly in Anglo-Saxon countries. These devices were a source of satisfaction for the participants and effective in terms of learning and acquisition of skills but they were mainly centered on one unique theme (e.g., recognition of oral pathologies) and were not focused on the orthodontic discipline [18][19]. However, an innovative complete distance continuing education environment could have many advantages, such as flexibility, lower costs, no office closing and accreditation by the French body of Continuing Professional Development (CPD) [20]–[24].

C. Existing distance learning device

According to the state-of-the-art [11]-[18], there was no complete distance learning environment adapted to the French orthodontic practitioners' needs. Only two complete websites dedicated to distance continuing education were intended for orthodontic practitioners: the World Federation of Orthodontists (WFO) and the *e-orthodontie.com* websites.

First, the WFO website, with online videoconferences access and its smartphone application (with notifications), is the most complete digital continuing education environment available to date, particularly concerning the diversified content, supports, and the scientific validity. Despite this, none of the interviewed practitioners were registered with WFO probably because this device was neither adapted (i) to their expectations and attitudes (ii) nor to their way of interacting with each other. Correlation between cultural and/or social dimensions with the use of a distance education device has already been highlighted in a previous study [25].

Second, the French *e-orthodontie.com* website has been created in 2007 without any prior user-centered research to assess practitioners' needs and expectations [19][20]. That could explain why this website was very little used by French orthodontic practitioners. This is evidenced by the fact that the activity in the forums section was close to zero.

D. The interactions within the devices

According the state-of-the-art [29], the loss of peer-to-peer interactions was the major drawback of the current distance education experiences for participants. That is why interaction represents one of the main issues to be considered for the design process. Nevertheless, several solutions are mentioned in the literature to create a kind of "e-presence" within the distance device, such as (i) virtual small groups of practitioners sharing same centers of interest or geographical proximity [27] (ii) creation of a collaborative e-portfolio [11][12] or (iii) tutoring [9][10]. But interactions between novices and their teachers *via* an e-portfolio were often limited, because, among other factors, teachers considered the digital feedback as a waste of time [11]. Concerning the remote tutoring, it remained generally underused because users struggled to meet their "ideal" tutoring model [9][10].

There are difficulties to maintain mutual commitment and trust in an online environment, hence the importance of examining the interactions within a current active CoP for creating a useful, usable, and acceptable digital education environment for orthodontic practitioners. We considered that an innovative distance continuing education environment, supported by the CoP members (and vice versa), could promote users' commitment. We based our approach on the horizontal social learning theories [3]-[6].

E. Contribution of a community of practice analysis for education device design

Several research-actions involving the design of training devices, in particular digital ones, are based on the notions of professional community in the education fields [33][34].

Nevertheless, CoPs case studies are very rare in the health sector and focus more on the education field (e.g., learning in a school context) [34]. One reason being that CoPs in the health sector (i.e., traditional and virtual) are fewer, less structured and often informal, therefore difficult to identify [35][36].

However, horizontal exchanges between peers represent an important source of cohesion and group identification within the CoP [1]–[6][21]. Besides, learning results from the interaction with other individuals and particularly with peers [3][4].

F. Definition of Community of Practice

First, the community of practice is defined by a common interest for a domain (this is what distinguishes CoP members from non-members) [41]. Second, a community, unlike project-focused teams, endures over time. The CoP members discuss, help each other, and share information in their fields. Third, they develop a shared repertoire of resources on their common practice: experiences, stories, tools, ways to handle issues. It is therefore by developing these three elements (domain, community, practice) that a CoP can grow and endure. The concept of communities of practice (CoP) is based on the idea that all individuals have always expanded their knowledge by discussing their practice with others, on a daily basis [42].

The social dimension of learning is therefore fundamental for understanding communities of practice [43]. Moreover, among the different types of communities (e.g., of interest, learners, epistemic) and working groups (e.g., project team, functional), the community of practice involves the strongest cohesions and interactions between its members. Regular interactions between peers within a CoP promote (i) mutual commitment, (ii) the emergence of a common project and (iii) the development of a shared repertoire. Exchanges have different values and purposes (e.g., learning, cohesion, identity affirmation). Some elements can either promote participation (e.g., trust, recognized expertise) or discourage it (e.g., fear of judgement, absence of answers, long delays). To last, a CoP must find a balance in interaction dynamics: nature, quantity, and rhythm.

Virtual learning communities have been on the rise in the past decade. Exchanges within these virtual communities are more fluid and faster than in traditional ones. However, they raise specific issues such as privacy, data security and the difficulty to create a user-friendly climate along with enabling confidence between the members [44].

G. Experts and novices

The Dreyfus model (Fig. 1) illustrates the five-phase trajectory from novice to expert including the intermediate stages (i.e., advanced beginner, competent and proficient) [45]. The process also involves several cognitive shifts ranging from the strict adherence to rules with lack of independent judgment, to deep contextualized intuitive understanding.

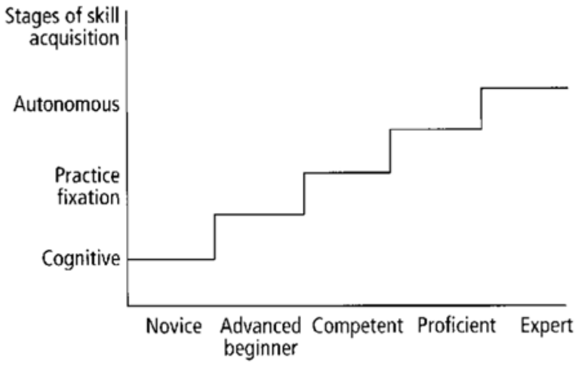


Figure 1. The 5 stages of learning within a CoP.

Learning paths vary depending on members' quest for personal and professional identity. The key focus is to find out their own way of becoming a "practitioner". This learning process is reflexive and built through access to heterogeneous practices and visions, which structure a real "landscape of practices" [39].

H. Community of practice and learning

The current remote videoconferences are based on the traditional vertical "teacher/learner" scheme. However, it is commonly accepted that a social learning environment is essential to foster adult education [46]. CoPs support their members' professional training by sharing their issues or experiences [47]. The resulting discussions from shared clinical experiences could help novices gain expertise by developing their ability to (i) analyze a new clinical situation, (ii) identify its consequences, (iii) adapt their behavior and (iv) consider other points of view [34][40][41]. Promoting peer interactions and encouraging them to share clinical issues are essential in theoretical and practical discipline such as orthodontics. This could help novices articulate these aspects [48].

III. Data & Methodology

To produce design recommendations on creating a useful, usable and acceptable digital education environment for orthodontic practitioners, three techniques have been used (Fig. 2). First, we observed the interactions and discursive activities within a virtual CoP "discutons entre spécialistes (let's discuss among specialists)". Second, we conducted four focus groups to identify behaviors related to continuing education, their needs, and expectations. Third, we carried out two surveys: one was addressed to members of a virtual CoP online and a second questionnaire survey was administrated during a congress, to collect data about attitudes and expectations towards continuing education. Figure 2 illustrates the adopted methodology and its objectives.

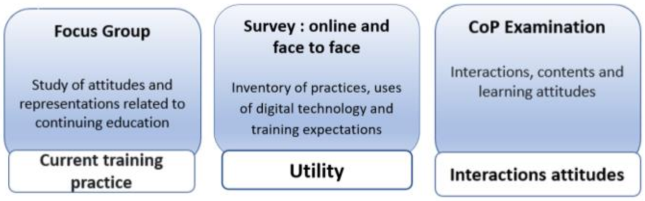

Focus Group
Study of attitudes and representations related to continuing education
Current training practice
Survey : online and face to face
Inventory of practices, uses of digital technology and training expectations
Utility
CoP Examination
Interactions, contents and learning attitudes
Interactions attitudes

Figure 2. The triangulation of our methodology.

A. Focus group

Four focus groups with 4 to 6 CoP novices were carried out: three focus groups were conducted before the health crisis and one after. The process was conducted in three stages:

(1) Identification of the difficulties, obstacles, and prospects of continuing education.

(2) Presentation of an existing French training system: the website *e-orthodontie.com*, to evaluate the participants' perception of digital training tools.

(3) Co-construction of "an ideal" website architecture dedicated to continuing education.

B. Questionnaire Survey (online and face-to-face)

The online survey was conducted among practitioners, members of a virtual CoP. The electronic survey was prepared and distributed by the software Limesurvey© to all CoP members, first on January 11, then on January 25, 2022 (n=59 CoP members, including 41 CoP experts and 18 novices).

This online survey was conducted to identify:

(1) Reasons for which practitioners became members.
(2) what the CoP actually provided for its members.
(3) The members status: novices or experts.

In this study, CoP novices were defined as either orthodontic resident (i.e., already qualified in dental medicine) or practitioner with less than three years of clinical experience. CoP experts were defined as orthodontic practitioners with more than three years of clinical experience.

The face-to-face survey was set up during a professional congress (n=42 practitioners, including 23 experts and 19 novices). The design of the questionnaire focused on the following items: current practices, digital uses, obstacles, success criteria of a training experience, opinions on distance continuing education and the vision of continuing education for the future.

C. Examination of a virtual CoP

The dual purpose of this examination was (i) an identification of the current interactions and (ii) description of the discursive activity (in term of content, nature of exchanges, objectives, rhythms, comments and likes generated, etc.) according to their experience (expert vs novice) within the CoP. This enables to study the discursive activity (e.g., rhythm, type of interactions, content within this CoP) and to identify the needs, attitudes, and expectations of the CoP members according to their experience (expert vs novice).

Qualitative discursive analysis enables the distinction between the different digital forms of interactions. Indeed, these revealed different level of participation, commitment and profiles of active members [49]:

- "Passive" digital participation includes document and posts reading
- "Active" participations, (i.e., contributions) includes:

o Non-verbal reactions (e.g., *like*s), specific to digital speech.

o Links, emojis, emoticons, which imply a higher level of participation (either consensus or controversy elements).

o Verbal contributions (e.g., publication of photographs, videos, or the writing of a publication, a question, an opinion). Their authors are considered as central and active CoP members.

On this basis, we examined the posted clinical cases (11 out of the 59 published in September 2021) and the peer comments and reactions. Qualitative analysis of peer comments was carried out using an existing quality grid (for content) [40] and on identified studies of digital discourse (for typology) [46] [47]

This qualitative analysis allowed us to (i) describe finely their forms and content (ii) edit some recommendations to encourage qualitative interactions that foster learning and members commitment.

D. Ergonomic inspection

The *eorthodontie.com* website (also presented during focus groups) was the subject of an ergonomic inspection, to determine if this interface was suitable for the practitioners [26]. We applied the ergonomic quality criteria developed by Bastien and Scapin. They organized the recommendations in the form of categories of ergonomic criteria such as: guidance, workload, brevity, explicit control, adaptability, error management, homogeneity/consistency, significance of codes and denominations, behaviors and compatibility [50]–[52]. This approach is based on the implicit idea that a digital device which meets these criteria is deemed adapted to the end user. The twin objective of this ergonomic inspection is to (i) understand better the reason why this website was little used by French orthodontic practitioners and therefore (ii) to edit some ergonomic recommendations toward the future design of an innovative device.

E. Data analysis

The focus group and the online survey data were analyzed as follows:

The textual analysis was carried out using free software IRAMUTEQ based on the R software and the Python language. After a manual thematic analysis, several automated analyzes were applied and in particular (i) the Reinert Descending Hierarchical Classification (DHC) model (ii) the Factorial Correspondences Analysis (FCA) and (iii) the similarity analysis. The DHC made it possible to divide the statements into classes marked by the contrast of their vocabulary. We completed DHC with a FCA which enabled us to observe the classes "geographical" proximity or distance. We also applied the similarity analysis when the number of segments was insufficient to obtain a saturation of the statements. We analyzed together the first three focus groups data (conducted before the health crisis), to compare them with the last focus group data (conducted after the health crisis). We also compared the online survey collected data between experts and novices (41 experts and 18 novices) to identify their common or divergent expectations and benefits of becoming member of a CoP.

The CoP posts and comments were analyzed as follows:

All posts and interactions (in the form of comments or likes) of the month of September 2021 were subjected to a thematic content analysis to group them within categories /themes. The nature of the exchanges (e.g., copresence, cooperation, collaboration, identification), correlated with different contents and levels of interaction, have been studied in accordance with Proulx's taxonomy [53]. Interactions level was measured as the sum of comments and/or likes of each publication (low: < or= to 10; medium: > to 10 and < or = to 20; and high: >20).

We analyzed the comments (i.e., categories, feedback type and specific application) generated by clinical case posts based on an evaluation grid of the "quality" of peer comments, produced in a previous study [5].

The face-to-face questionnaire collected data were analyzed as follows:

R studio analysis software was used for all statistical analyses. A *p*-value of less than 0.05 was chosen as the minimum significance value.

Two univariate analyzes were carried out according to age and years of experience.

The Welch test made it possible to evaluate the covariance between several quantitative and qualitative variables with several modalities. For example, the number of days devoted to continuing education, between several groups (i.e., novices and experts).

The Khi2 test provides the means to compare the distributions of a categorical variable (e.g., obstacles to following a training course) between several groups (e.g., novices and experts). When the Khi2 application conditions could not be met, a Fisher's test was performed.

IV. MAIN RESULTS

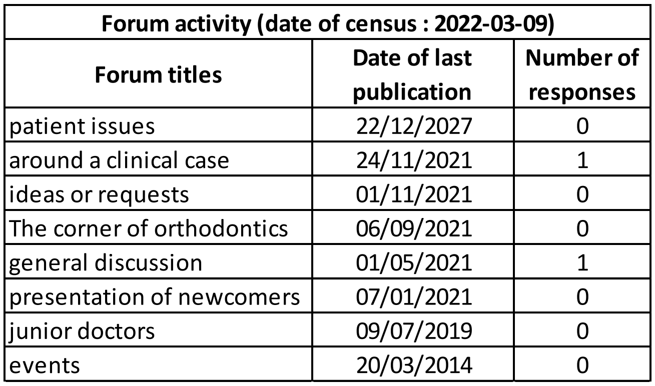
A. *Contribution of the ergonomic inspection*

If the main objective of this website is to train practitioners based on interactivity (the site presentation specifies: *"interactive site of orthodontics, presentation of clinical cases, photos, videos and medical forums. Orthodontic training guaranteed!*"), it is struggling to meet its stated goal.

A descriptive approach made it possible to measure the current activity of the *eorthodontie.com* website, by relying on the "forum", "articles" and "downloads" sections. The census of "forum" activity (Fig. 5) revealed that the publication activity was sporadic and generated very few interactions (compared to the Facebook© discussion group "Let's discuss among specialists"). The *eorthodontie.com* website also seemed more active with patients than with intended end-users (i.e., orthodontics practitioners). Indeed, the most recent discussions in the "forums" section mainly concerned patients undergoing orthodontic treatment who were seeking another opinion.

The interaction, therefore, appeared to be very limited because:

- the only answers came from a single moderator,
- there was no discussion between practitioners,
- the last posts dated more than three months ago.



Forum activity (date of census : 2022-03-09)		
Forum titles	Date of last publication	Number of responses
patient issues	22/12/2027	0
around a clinical case	24/11/2021	1
ideas or requests	01/11/2021	0
The corner of orthodontics	06/09/2021	0
general discussion	01/05/2021	1
presentation of newcomers	07/01/2021	0
junior doctors	09/07/2019	0
events	20/03/2014	0

Figure 3. Table summarizing the activity (date of last publication and responses) of the available discussion groups.

The ergonomic criteria inspection revealed that:

- Overall informational density was too high
- Immediate feedback was well respected
- Workload was increased due to the lack of brevity and conciseness and the high number of unnecessary actions
- Certain procedures were unnecessarily too heterogeneous
- Readability could be improved

The interface adapted very little to the user's experience and there was no error management. Regarding the content, it lacked completeness, updating, and was not sufficiently scientifically validated. This website examination emphasized the importance of (i) respecting the ergonomic criteria during the device design and (ii) conducting a preliminary practitioner needs assessment.

B. *Need for "e-presence"*

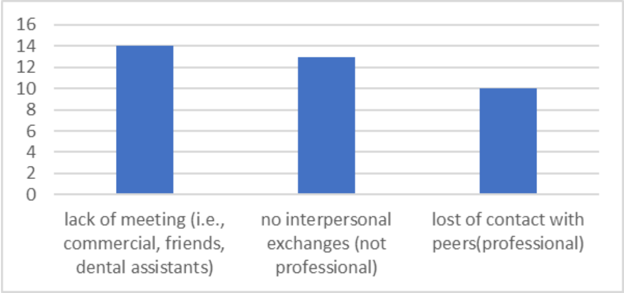


Figure 4. Main disadvantages of distance learning mentioned by practitioners in the questionnaire survey (number of occurrences in descending order after manual thematic grouping).

According to the state-of-the-art, the loss of contact with peers as well as the lack of meetings and exchanges were the main distance learning disadvantage for CoP novices and experts interviewed (questionnaire survey collected data) [54] (Fig. 3). The Welch test highlighted that for younger practitioners (Fig. 4) the decision to participate in a face-to-face congress was correlated with the presence of a close member (*p-value*: 5.366e-09), probably because of their peripheral position within the CoP.

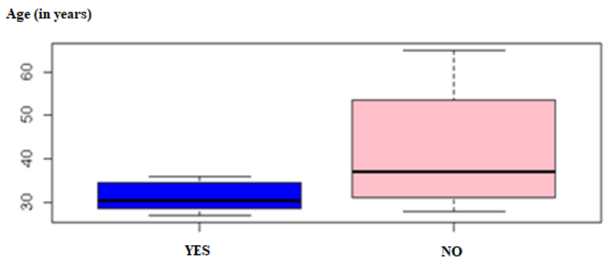


Figure 5. Boxsplot representing the importance of the presence of a close person to attend a training according to the age of the participants.

These results confirmed the importance to consider the problem of remote presence in the device design.

C. *Contribution of the virtual CoP discursive analysis*

This innovative continuing education environment is addressed to French orthodontic practitioners who are mostly already involved in an informal virtual CoP, built on Facebook© in 2014. This virtual active CoP "*let's discuss among specialists"* (in French: discutons entre spécialistes) has significantly grown these last years. The growth of the informal virtual CoP these last three years (see Fig. 2) seemed to be an underlying trend (i.e., +170% members since 2019). Indeed, the first COVID-19 lockdown (i.e., start date 03/17/2020) did not seem to have modified this growth. In 2022 February, this CoP was gathering 1082 practitioners, representing almost half of the population (i.e., 2420 orthodontic specialists).

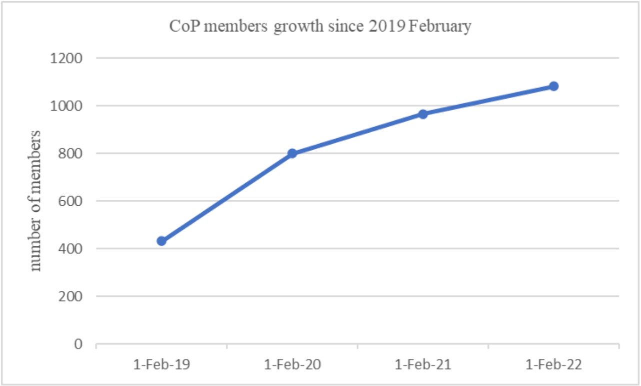


Figure 6. CoP growth since 2019.

The analysis of the publication's rhythm in September 2021 (n=59) revealed its cyclical aspect (see Fig. 7). The analysis of the authors' status showed that the start of a new cycle of publications coincided with a publication by a central CoP member (i.e., moderator, administrator, or recognized expert): their role was crucial in maintaining and developing the interaction

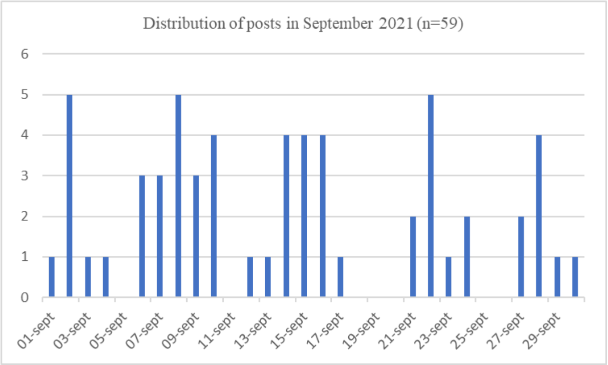
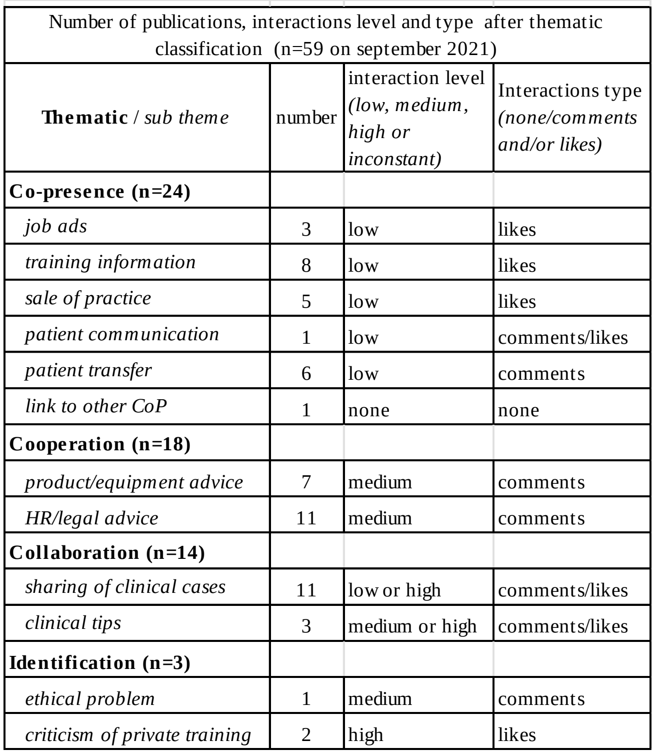


Figure 7. Distribution of the posts in September 2021.

.

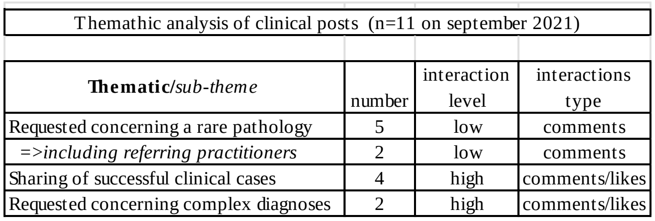


Number of publications, interactions level and type after thematic classification (n=59 on september 2021)			
Thematic / *sub theme*	number	interaction level *(low, medium, high or inconstant)*	Interactions type *(none/comments and/or likes)*
Co-presence (n=24)			
job ads	3	low	likes
training information	8	low	likes
sale of practice	5	low	likes
patient communication	1	low	comments/likes
patient transfer	6	low	comments
link to other CoP	1	none	none
Cooperation (n=18)			
product/equipment advice	7	medium	comments
HR/legal advice	11	medium	comments
Collaboration (n=14)			
sharing of clinical cases	11	low or high	comments/likes
clinical tips	3	medium or high	comments/likes
Identification (n=3)			
ethical problem	1	medium	comments
criticism of private training	2	high	likes

Figure 8. Publications thematic analysis and the level of interactions generated.

Figure 8 shows the publications thematic analysis and the level of interactions generated. Most publications were of the order of co-presence among members, creating few reactions (mostly *likes)*. Their content were mainly informational. Publications on the mode of cooperation were less frequent but generated a higher level of interaction (mostly *comments*). The collaborative publications, generating a sustained interaction (i.e., clinical cases and clinical tips) were also rarer. During the month of September, three publications with strong identity value were published (i.e., one ethical problem and two criticisms of private training). These elicited many reactions (*likes* or *comments*).

However, all CoP members did not publish in all categories. The publications allowing either reflection on the orthodontics practice or collaboration among members, came exclusively from the CoP core experts, administrators, and moderators. The novices never participated in the form of posts or comments and very rarely in the form of likes. This observation is consistent with the data collected by focus group: all CoP novices (pre and post COVID-19 focus group) expressed their fear of being judged by the CoP experts. That was indeed the main barrier to their participation [5][37]. It is for this reason that anonymity was such a strong novices' expectation.



Themathic analysis of clinical posts (n=11 on september 2021)

Thematic/*sub-theme*	number	interaction level	interactions type
Requested concerning a rare pathology	5	low	comments
=>*including referring practitioners*	2	low	comments
Sharing of successful clinical cases	4	high	comments/likes
Requested concerning complex diagnoses	2	high	comments/likes

Figure 9. Shared clinical cases detailed thematic analysis.

Figure 9 shows that practitioners never shared failures or treatments incidents, although this was an explicit strong request from novices, according to post COVID-19 focus group collected data.

The analysis of the 11 clinical contributions allowed us to carry out thematic groupings. The peer reactions and comments were correlated to the topic.

- Four clinical contributions were "well-ended" clinical cases (treated by an innovative technique). These generated a variable number of comments, mostly in the form of likes, but also, thanks, encouragements, or requests to use the same technique. They were published only by the active and recognized CoP experts.
- Two contributions were requests for help with a complex diagnosis or treatment plan. The peers' comments were numerous, and their form diversified: link to videos, photographs or links to other published cases, articles, *etc.* They sometimes gave rise to (i) debates (between active CoP experts only), (ii) searches for a consensus (iii) discussions on corollary subjects (e.g., techniques, devices) (iv) expressions of support toward peers (i.e., the author or other practitioners)
- Two were requests regarding regional issues (e.g., search for a genetic reference center near the practitioner), generating few reactions (i.e., likes) and comments.
- Three publications about rare clinical issues (rare pathologies or technical complex situations). This category generated few comments, in the form of sharing clinical experience, purely informational (i.e., no request)

The description of the peer comments was carried out using the evaluation grid edited by Ortoleva & Bétrancourt (2016) and allowed us to make the following observations:

• The 11 clinical cases were exclusively posted by active CoP experts.

Among these 11 clinical cases, none could be considered as "failed" or "treatment incident". However, the novices interviewed in the focus groups clearly expressed their expectation of publishing treatment incidents as well. According to the literature, sharing these failures represents an excellent source of learning [5].

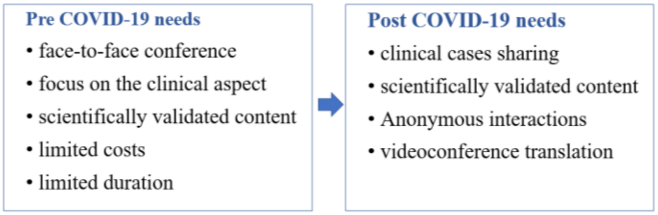
• None of the peers' comments were personal clinical situations, they referred only to "imaginary" situations or previously published clinical cases.

• As of the two complex clinical cases, only experts discussed the best way to handle the situation. This is the category in which we observed the most (i) discursive precautions, (i.e., politeness) and (ii) diversified supports (e.g., link to other clinical cases, videos, training).

• From a formal point of view, concerning the readability, intelligibility and exhaustiveness of the comments, these criteria were well respected.

D. Impact of the COVID-19 crisis on the CoP members learning needs

The comparison between the focus groups data collected before *versus* after the health crisis enabled us to describe finely the changes of continuing education perception, raised by the literature [15]–[17]. Regarding the interactions, in the pre COVID-19 focus group, the lack of informal exchanges between peers was a significant barrier to distance learning. The "ideal" learning experience was a face-to-face conference, with limited costs and duration. In contrast, in the post COVID-19 focus group, the" ideal" learning experience consisted in clinical cases sharing (i.e., especially failed treatment) illustrated step by step, anonymous, internet-based literature search, scientifically validated content, and videoconference instant translation into French. The need to translate was strong for CoP novices, probably because they were afraid of misunderstandings without being able to detect them. The health crisis changed deeply the practitioners' perception toward distance learning (Fig. 10). According to the literature, an innovative complete distance continuing education environment could henceforth meet many CoP members' needs [21][22][23].



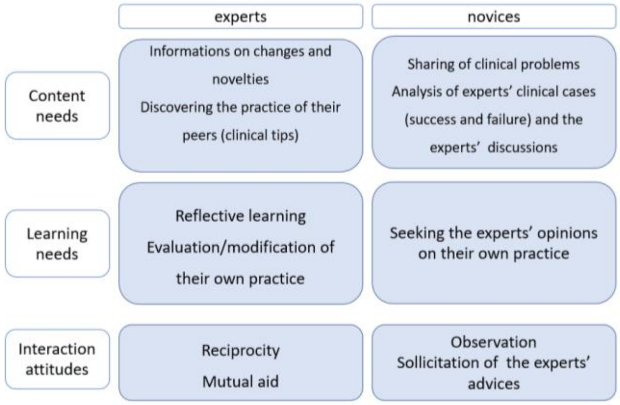
Pre COVID-19 needs
• face-to-face conference
• focus on the clinical aspect
• scientifically validated content
• limited costs
• limited duration

Post COVID-19 needs
• clinical cases sharing
• scientifically validated content
• Anonymous interactions
• videoconference translation

Figure 10. Evolution of experts and novices' needs after the COVID-19 crisis

E. Experts/novices: interactions attitudes, needs, and expectations

The distinct similarity analysis produced from novices and experts' responses to the online survey, allowed us to distinguish their expectations and needs towards the CoP. Figure 11 shows two different profiles in terms of content, interaction needs and attitudes within the virtual CoP "*let's discuss among specialists"*. The experts expected to (i) be informed about the novelties, (ii) discover the practice and clinical tips of their peers. Their main goals were to evaluate their own practice and eventually modify it: that was a reflective learning process based on reciprocity. Concerning novices' needs, they expected to obtain expert opinions and were in an observant attitude.



	experts	novices
Content needs	Informations on changes and novelties Discovering the practice of their peers (clinical tips)	Sharing of clinical problems Analysis of experts' clinical cases (success and failure) and the experts' discussions
Learning needs	Reflective learning Evaluation/modification of their own practice	Seeking the experts' opinions on their own practice
Interaction attitudes	Reciprocity Mutual aid	Observation Sollicitation of the experts' advices

Figure 11. Experts and novices' needs in terms of content, learning and interaction.

F. Interest of being member of a CoP

According to the online survey collected data, the reasons why practitioners become members of a virtual CoP depended on their status.

For the novices, the members of a virtual CoP was mainly looking for a "network" and "professional announcements". The verbs "to share", "to obtain opinions" and "to see" which accompanied the notion of " clinical cases". For the novices, the added value of being a part of a virtual CoP was "to see others' clinical cases" and "to discover their peers' practice". This highlights to the importance of interactions and sharing practice for their professional development. These results were consistent with the observational posture of novices within a CoP.

As of the experts, the members of a virtual community of practice were mostly looking for "what's new", "tips and tricks" and to "help each other". The recurring verbs associated with the notion of "clinical cases" were "to learn" and "to exchange".

Concerning this group, the interactions take place based on reciprocity, which is consistent with their more central position within the CoP.

G. Interaction quality improvement

To improve the efficiency of the practitioners' comments in terms of learning and collaboration, a participation charter could be draw up, according to the peers' comments quality evaluation grid [32]:

- By specifying the rules of participation (e.g., use of a friendly tone), this charter could encourage the participation of novices who were afraid to be judged by the CoP experts.
- It could improve the "quality" of peer discussions by encouraging them to (i) share personal clinical cases, (ii) come up with questions, (iii) make suggestions, (iv) share scientific articles. That could indeed promote learning and address the needs of novices.

The publication of recommendations for speakers could allow videoconferences to meet (i) the expectations of CoP members, (ii) and comply with recommendations for good practice. These suggestions could encourage them:

- To cite scientifically validated articles.
- To be exhaustive and up-to-date.
- To focus on daily practice and clinical aspects.
- To answer questions from practitioners.
- To produce a bibliography to deepen the subject.

H. Several requirements to promote peers' interactions

The current virtual CoP via Facebook© did not allow to create small discussion groups, nor to publish anonymously. An innovative continuing education environment should offer these possibilities to encourage novices' participation and ultimately stimulate interactions between peers. According to the state-of-the-art and our collected data, some criteria should be respected to promote practitioners' participation:

First, participation in discussions forums could be done under a pseudonym. However, each practitioner's status should be known (novices/experts), so that novices could trust in the posted information.

Second, the content scientific validity could be ensured by various means:
-Review by known International/European clinical experts.
-Review by teachers from universities.
-Review by a mixed college (universities teachers and clinical experts).

Some of these experts should also be involved in facilitating the forum and take on the role of moderator to promote the interactions, as in the virtual CoP "*let's discuss among specialists".*

Third, the device should offer the possibility of exchanging on his/her clinical cases via a forum, seeking the opinions of other practitioners or even having access to very detailed clinical cases (step by step).

Fourth, the device should allow the creation of limited or extended discussions groups based on professional status (expert/novices). The geographical discussion group could also be relevant according to the CoP discursive analysis: 2 of the 11 clinical posts were indeed requests for referring practitioners in the same region (see Fig. 5).

Finally, active (e.g., contributions, likes, emojis) and passive (e.g., reading content) participations are valuable to build trust between members. In the Facebook virtual community, passive participation is not known, yet passive participation is essential for authors to be aware of being read. Thus, to value the passive participation, the number of views for each post could be visible

V. Discussion

The collected data (focus group, online survey, virtual CoP examination) agreed and complemented each other. This confirmed the interest of adopting a data triangulation method to formulate relevant recommendations [56].

Although learning within a CoP is a trajectory from novice to expert passing through intermediate stages, the data analysis conducted by dividing them into two groups (novices vs experts) allowed us to reveal different attitudes, needs and expectations in terms of continuing education.

It is commonly accepted that novices participated less than experts, because of their peripheral position within the CoP [3]-[6]. However, an education device should encourage all CoP members to participate on a voluntary basis, to reduce the feeling of loneliness and foster their commitment [4]. But, if virtual CoPs share the same principles than traditional ones (e.g., commitment and mutual trust), this is more difficult to maintain in an online environment [57].

Our collected data explained more precisely why the WFO and the *e-orthodontie.com* websites did not match users' expectations. Concerning the WFO website, there was a strong language barrier: in all focus group, the need to translate everything into French was commonly shared. Concerning the French *e-orthodontie.com* website, the content was perceived as not scientifically valid by interviewed practitioners. Moreover, this website was accessible to patients, specialist, and non-specialist orthodontic practitioners. This "open access" was the subject of numerous criticisms by all the interviewed practitioners. In addition, these websites did not qualify for Continuing Professional Development credits.

In addition, the virtual CoP "let's discuss among specialists", although active and growing, is struggling to involve novices despite expressing a significant need to share their clinical cases and their "failures". Furthermore, the impossibility of forming small discussion groups and the lack of anonymity seem to hinder their participation. However, this sharing of clinical cases on the part of novices and the solicitation of peer comments on them could help them better articulate the practical and theoretical dimensions of the discipline. Moreover, the clinical cases presented by the experts are above all very complex diagnoses (never failures, etc.) or successful clinical cases but whose treatment management deviates somewhat from the recommendations.

All surveys revealed indeed the significant tension within this CoP related to the various academic backgrounds (specialists versus non-specialists). The open or limited access of non-specialists to the innovative distance learning environment should be carefully considered: the specialists considered the non-specialists as an outgroup of the CoP, whereas the non-specialist probably considered the specialists as experts of the CoP.

This paper showed that orthodontic practitioners commonly needed (i) scientifically validated content, (ii) extensive discussion and limited groups, (iii) anonymous, (iv) publications on clinical cases (successful AND unsuccessful). These results were consistent with the state-of-the-art. But contrary to the literature, in our study, the discussion forums group should be centered on the professional status (CoP novices and/or experts) and not on the center of interest [27].

It would have been interesting to carry out focus groups of CoP experts, but professional constraints (solitary practice, geographically scattered, lack of time) prevented us from doing so. Nevertheless, the online survey enabled us to include mostly CoP experts. The experts were numerous either because they participated more actively into the CoP, and/or because they were more represented there.

We conducted this user-centered psycho-ergonomic study by limiting the notion of users to practitioners who need to be trained and not to speakers and/or trainers/facilitators/tutors who provide resources and/or animate the CoP. However, it would seem necessary to identify the needs of all the actors to promote the acceptability of the system. It would therefore be interesting to extend this work by evaluating the different stages of design by the different actors.

VI. Conclusion and Future Work

A complete, careful analysis of the orthodontic practitioners' needs, expectations, and interactions behavior within the virtual active CoP *"let's discuss among specialists"* was done for this innovative distance environment to comply with the criteria of usability and acceptability.

According to our data collection, a comprehensive distance learning environment could meet many novices and experts' expectations. Indeed, the CoP novices reported their need to (i) interact with experts anonymously (to avoid being judged), (ii) create restricted or extended online discussion, (iii) ask for questions about all available content (e.g., videoconferences, articles), and (iv) be informed of news by notification. The needs and attitudes of novices and experts we described in this study are supported by the data on the CoPs [1]-[6], particularly concerning cohesion, sharing of experiences and identity needs. However, the way to proceed is specific to each profession and, to our knowledge, no previous study has analyzed the orthodontic practitioners' community.

This research revealed that discussions on the posted clinical experiences constituted the CoP added value perceived by its members and helped novices articulate the theoretical and practical dimensions of orthodontics. As such, the sharing of clinical experience must be encouraged in the future system

This study allowed us to identify the CoP members needs and expectations in terms of (i) content (and the categories structuring it), (ii) expected interactions between novices or experts (e.g., rhythm, themes, anonymity, etc), (iii) scientific validity, (iv) sharing or observing the peers' positives or negatives clinical experiences. Our findings indicates that (i) COVID-19 crisis modified the CoP members learning needs and (ii) the interaction needs, attitudes, and expectations of CoP novices and experts were different. On this basis, several requirements in term of interactions and contents have been proposed.

This users' center research showed that an innovative education environment would greatly enrich the CoP, particularly in terms of content, support, and variety of possible exchanges. All focus groups participants co-created a website architecture and discussed their expectations in terms of supports and contents to design an "ideal" distance learning device. The contents and supports will be the focus of a future article.

Our user-centered approach must be extended during the design/redesign phases by empirical methodology at different stages without and /or with "real" users, to ensure compliance with the device ergonomic criteria [58]. In the next phases, the concept of users should encompass lecturers and facilitators.

The security and legal standing of shared medical data such as X-rays and/or photographs of patients' needs to be addressed. Further studies on the security aspects of the device are also important to be conducted to minimize the risks of malicious attacks and gain more confidence from the practitioners.

Further experimentation should be conducted, including more in-depth investigation of practitioners' expectations during the post COVID-19 period to justify usefulness of the proposed requirements.

References

[1] A. Mailloux and J. Dinet, "Requirements Analysis Towards Future Design of an Innovative Distance Learning Device Intended for French Orthodontic Practitioners Contribution of a Community of Practice Analysis," The Sixteenth International Conference on Digital Society, June 2022.

[2] A. Jezegou, "distance in formation. First milestone for an operationalization of the theory of transactional distance". Distances et savoirs, Vol. 5, no. 3, pp. 341-366, 2007.

[3] J.-H. Park and H.-J. Choi, "Factors Influencing Adult Learners' Decision to Drop Out or Persist in Online Learning," Learning & Technology Library (LearnTechLib)" 2009. https://www.learntechlib.org/p/74987/ [retrieved: September 2022]

[4] J. Lave and E. Wenger, "Situated learning: Legitimate peripheral participation," New York, NY, US: Cambridge University Press, pp. 26-49, 1991.

[5] G. Ortoleva and M. Bétrancourt, "Supporting productive collaboration in a computer-supported instructional activity: peer feedback on critical incidents in health care education," J. Vocat. Educ. Train., vol. 68, no. 2, pp.178-197, Apr. 2016.

[6] M. Gosselin, A. Viau-Guay, and B. Bourassa, "The different learning processes experienced by health professionals participating in a community of practice," Phronesis, vol. 6, no 3, pp. 36-50, 2017.

[7] V. Glikman, "Learners and tutors: a European approach to human mediations," Education permanente, vol 152, pp. 55-69, 2002. [retrieved: May 2022]

[8] V. Glikman, "Remote Tutor," in Distance education tutoring, B. D. Lievre, C. Depover, A. Jaillet, D. Peraya, and J.-J. Quintin, De Boeck, pp. 137-158, 2011.

[9] C. Vernazza, "Introduction of an e-portfolio in clinical dentistry: staff and student views," vol. 15, no. 1, pp. 36-41, Eur J Dent Educ, 2011.

[10] R. L. Kardos, J. M. Cook, R. J. Butson, and T. B. Kardos, "The development of an ePortfolio for life-long reflective learning and auditable professional certification," Eur. J. Dent. Educ. Off. J. Assoc. Dent. Educ. Eur., vol. 13, no. 3, pp. 135-141, August 2009.

[11] K. Errabi, "Demand-Pull" or "Technology-Push": "survey of recent literature and new econometric tests," p. 165, 2017.

[12] A. Chaptal, "Abstract," Distances Savoirs, vol. 1, no. 1, pp. 121-147, 2003.

[13] S. Caro Dambreville, "Design of digital documents. Methodological path," *Doc. Numér.*, vol. 12, no. 2, pp. 7-22, 2009.

[14] G.-L. Baron and E. Bruillard, "Informatics and its users in education," Paris: PUF, 1996.

[15] X. Liu, J. Zhou, L. Chen, Y. Yang, and J. Tan, "Impact of COVID-19 epidemic on live online dental continuing education," Eur. J. Dent. Educ., vol. 24, no. 4, pp. 786-789, Nov. 2020.

[16] H. C. Cheng, S.-L. Lu, Y. C. Yen, P. Siewchaisakul, A. M. F. Yen, and S. L. S. Chen, "Dental education changed by COVID-19: Student's perceptions and attitudes," BMC Med. Educ., vol. 21, no. 1, p. 364, July 2021.

[17] R. Elledge, R. Williams, C. Fowell, and J. Green, "Maxillofacial education in the time of COVID-19: the West Midlands experience," Br. J. Oral Maxillofac. Surg., vol. 60 no. 1, pp. 52-57, July 2020.

[18] L. Browne, S. Mehra, R. Rattan, and G. Thomas, "Comparing lecture and e-learning as pedagogies for new and experienced professionals in dentistry," Br. Dent. J., vol. 197, no. 2, pp. 95-97, July 2004

[19] L. Stow and D. Higgins, "Development and evaluation of online education to increase the forensic relevance of oral health records," Aust. Dent. J., vol. 63, no. 1, pp. 81-93, March 2018.

[20] K. M. Bernie, E. T. Couch, and M. Walsh, "Perceptions of California Dental Hygienists Regarding Mandatory Continued Competence Requirements as a Condition of License Renewal," J. Dent. Hyg. JDH, vol. 90, no. 5, pp. 275-282, Oct. 2016.

[21] M. Bonabi, S. Z. Mohebbi, E. A. Martinez-Mier, T. P. Thyvalikakath, and M. R. Khami, "Effectiveness of smart phone application use as continuing medical education method in pediatric oral health care: a randomized trial" BMC Med. Educ., vol. 19, no. 1, p. 431, Nov. 2019.

[22] K. P. Klein, K. T. Miller, M. W. Brown, and W. R. Proffit, "In-office distance learning for practitioners," Am. J. Orthod. Dentofac. Orthop. Off. Publ. Am. Assoc. Orthod. Its Const. Soc. Am. Board Orthod., vol. 140, no. 1, pp. 126-132, Jul. 2011.

[23] E. W. Odell, C. A. Francis, K. A. Eaton, P. A. Reynolds, and R. D. Mason, "A study of videoconferencing for postgraduate continuing education in dentistry in the UK--

the teachers' view," Eur. J. Dent. Educ. Off. J. Assoc. Dent. Educ. Eur., vol. 5, no. 3, pp .113-119, Aug. 2001.

[24] N. Mattheos et al., "Potential of information technology in dental education," Eur. J. Dent. Educ., vol. 12, pp. 85-92, Feb. 2008.

[25] K. T. Miller, W. M. Hannum, T. Morley, and W. R. Proffit, "Use of recorded interactive seminars in orthodontic distance education,"Am. J. Orthod. Dentofac. Orthop. Off. Publ. Am. Assoc. Orthod. Its Const. Soc. Am. Board Orthod., vol. 132, no 3, pp. 408-414, sept. 2007.

[26] J. M. C. Bastien and D. L. Scapin, "A validation of ergonomic criteria for the evaluation of human-computer interfaces," Int. J. Human– Computer Interact., vol. 4, no. 2, pp. 183-196, Apr. 1992.

[27] É. Brangier and J. M. C. Bastien. "The evolution of the ergonomics of computer products: accessibility, usability, emotionality and influenceability," Presses Universitaires de France, pp. 307-328, 2010. [retrieved: Apr. 2022].

[28] P. Rekawek, P. Rice, and N. Panchal, "The impact of COVID-19: Considerations for future dental conferences," J. Dent. Educ., vol. 84, no. 11, pp. 1188-1191, Nov. 2020.

[29] M. Grangeat, "L. S. Vygotski: group learning", Éditions Sciences Humaines, pp. 134-141, 2016

[30] H. Spallek et al., "Supporting emerging disciplines with e-communities: needs and benefits," J. Med. Internet Res., vol. 10, no. 2, Apr-Jun. 2008.

[31] P. Ratka-Krüger, J. P. Wölber, J. Blank, K. Holst, I. Hörmeyer, and E. Vögele, "MasterOnline periodontology and implant therapy-revisited after seven years: A case study of the structures and outcomes in a blended learning CPD," Eur. J. Dent. Educ. Off. J. Assoc. Dent. Educ. Eur., vol. 22, no. 1, pp. 7-13, Feb. 2018.

[32] L. Cifuentes, G. Maxwell, and Ș. Bulu, "Technology Integration Through Professional Learning Community," J. Educ. Comput. Res., vol. 44, pp. 59-82, Jan. 2011.

[33] S. Ouellet, I. Caya, and M. P. Tremblay, "The contribution of a learning community to developing collaborative and inclusive practices: action research," Éducation Francoph., vol. 39, no. 2, pp. 207-226, 2011.

[34] A. le May, "Communities of Practice in Health and Social Care," John Wiley & Sons, p. 144, 2009.

[35] L. C. Li, J. M. Grimshaw, C. Nielsen, M. Judd, P. C. Coyte, and I. D. Graham, "Use of communities of practice in business and health care sectors: A systematic review," Implement. Sci., vol. 4, no. 1, p. 27, Dec. 2009.

[36] M. Poirier, "Communities of Practice in Mental Health," vol. 3, no. 1, 2010. https://cremis.ca/publications/articles-et-medias/les-communautes-de pratiques-en-sante-mentale/ [retrieved: September 2022]

[37] E. Wenger, "Communities of Practice and Social Learning Systems: the Career of a Concept," in Social Learning Systems and Communities of Practice, C. Blackmore, London: Springer, pp. 179-198, 1998.

[38] E. Wenger, R. A. McDermott, and W. Snyder, "Cultivating Communities of Practice," Boston, Mass: Harvard Business Review Press, 2002.

[39] P. Lièvre, E. Bonnet, and N. Laroche, XXI. Etienne Wenger, "Community of practice and social theory of learning," EMS Editions, pp. 427-447, 2016.

[40] J. Brown, "Online learning. CDS to present debut webinar on Direct Resin Pearls," CDS Rev., vol. 103, no. 5, pp. 14-15, Oct. 2010.

[41] I. Bourdon, N. Teissier, and C. Kimble, "Relationships and participation within a virtual community of practice: case study in a multinational engineering company", Recherches en sciences de gestion, vol.100, pp. 121-141, 2014.

[42] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change", Psychol. Rev., vol. 84, no. 2, pp. 191-215, 1977

[43] A. Amiel, J.-F. Camps, G. Lutz, F. Plégat-Soutjis, and A. Tricot, "Designing an Open and Distance Learning System", p. 24, 2002.

[44] S. E. Dreyfus, "The Five-Stage Model of Adult Skill Acquisition," Bull. Sci. Technol. Soc., vol. 24, no. 3, pp. 177-181, 2004.

[45 L. Stoll, R. Bolam, A. McMahon, M. Wallace, and S. Thomas, "Professional Learning Communities: A Review of the Literature," J. Educ. Change, vol. 7, no. 4, pp. 221-258, 2006.

[46] M. C. Smith and T. Pourchot, "Adult Learning and Development: Perspectives From Educational Psychology," Routledge, p. 296, 2013.

[47] G. Ortoleva, M. Bétrancourt, and S. Morand,"Between personalization and collective constraints: A user-centric approach for the implementation of a digital Pedagogical Monitoring booklet," Sci. Technol. Inf. Commun. Pour LÉducation Form., vol. 19, no. 1, p. 233-251, 2012,

[48] M. Alomran, "Transdisciplinary qualitative approach to analyze interactions and instrumentations in online discussions," Comprehensive research & qualitative approaches (the 10th Doctoriades), Oct. 2021. [retrieved: Juil. 2022]

[49] J. M. C. Bastien and D. L. Scapin, "A validation of ergonomic criteria for the evaluation of human-computer interfaces," Int. J. Human– Computer Interact., vol. 4, no. 2, pp. 183-196, Apr. 1992.

[50] J. M. C. Bastien, "Ergonomic criteria: a step towards a methodological aid to the evaluation of interactive systems," Doctoral Thesis, Paris 5, 1996.

[51] C. Bastien et D. Scapin, " User-centered interactive software design: steps and methods," in Ergonomie, Presses Universitaires de France, pp. 451-462, 2004.

[52] S. Proulx, "Virtual communities: what makes a connection," in S. Proulx, L. Poissant and M. Sénécal, "Virtual communities: thinking and acting in a network," Presses de l'Université Laval, pp. 13-26, 2006.

[53]] H. Spallek et al., "Supporting emerging disciplines with e-communities: needs and benefits", J. Med. Internet Res., vol. 10, no. 2, Apr-Jun. 2008

[54] L. Arcand, "The community of practice a relevant tool: summary of knowledge adapted to the context of public health," INSPQ, Nov. 2017 [retrieved: Dec 2022]

[55] T. Apostolidis, "Social representations and triangulation: theoretical methodological issues," In: J. C. Abric, "Methods of studying social representations", Ed Érès, pp. 13-35, 2003.

[56] B. Mercieca, "What Is a Community of Practice?," in Communities of Practice: Facilitating Social Learning in Higher Education, J. McDonald and A. Cater-Steel, Singapore: Springer, pp. 3-25, 2017.

[57] C. Lewis and J. Gould, "Designing for usability: key principles and what designers think," Communications of the ACM , pp. 300-311, 1985. [retrieved: Mar 2022]

Explainable Kinship: A Broader View on the Importance of Facial Features in Kinship Recognition

Britt van Leeuwen[a], Arwin Gansekoele[b], Joris Pries[c], Etienne van de Bijl[d] and Jan Klein[e]

Centrum Wiskunde & Informatica, Stochastics group
Science Park 123, Amsterdam, the Netherlands
Email: [a]britt.van.leeuwen@cwi.nl, [b]arwin.gansekoele@cwi.nl,
[c]jorispries@gmail.com, [d]etienne.van.de.bijl@cwi.nl,
[e]jan_g_klein@outlook.com

***Abstract*—Kinship Recognition, the ability to distinguish between close genetic kin and non-kin, could be of great help in society and safety matters. Previous studies on *human* kinship recognition found interesting insights when looking for the most important features. Results showed that analyzing only the top half of a face gives equal or even better performance compared to analyzing the whole face. In this paper, we aim to find the important features for *automated* kinship recognition based on the theory of *human* kinship recognition; this set of features was researched using features from pre-trained metrics from the StyleGAN2 model. Three different experiments were performed focusing on different aspects of facial features. We found that the most important facial features from the selection of 40 features are mostly focused on the facial hair traits. Furthermore, age-related features were found to be very important. This set of features does not entirely comply with the set of features important in *human* kinship recognition. Previous research has shown *human* kinship recognition performance does not decrease when removing the bottom half of the image of the face. In contrast, our results show that for *automated* kinship recognition, removing either the bottom or the top half of a face results in a decrease in the performance of our classifiers. Moreover, only using a selection of facial features corresponding with the important features in human kinship recognition did not prove to be sufficient for the task of Kinship Recognition.**

***Keywords*—kinship recognition; StyleGAN2; Families-in-the-Wild; feature importance; transfer learning.**

I. Introduction

This work is an extension on our previous research in [1] on the importance of facial features in Kinship Recognition.

A. Kinship Recognition

One of the fields in artificial intelligence that is currently of great interest is computer vision. Computer vision is defined as the study domain that revolves around techniques developed to automate seeing and understanding the contents of digital images such as photographs and videos by computers [2]. This new field started to emerge around the 1960s [3]. However, projects such as getting a computer to describe what it saw via a linked camera, proved much more complex than first thought [4]. Computer vision began to rise after a couple of decades, as the internet advanced and, therefore, access to data improved. At that time (the 80s and 90s), it was facial recognition that grew to be more promising. Subsequently, with the boost of the internet and following later social media, we came as far as Facebook using face recognition every day [5]. One of the subjects that keeps computer vision attractive today is face recognition. From unlocking your phone using your face to automated passport checking at an airport, face recognition is used more often than we realize.

The subject of kinship recognition (KR) is fairly new within face recognition. Kinship recognition is the ability to distinguish between close genetic kin and non-kin. The distinction involves people who are directly related and people who are not. One example of the usage of KR is of families who are spread throughout multiple refugee camps. One of these cases involved a father and his daughter being in one camp, while his wife and other children were in another camp. It took them over a year to get reunited by the Red Cross Restoring Family Links [6]. Even with an organization on these reunion cases, it still takes them a considerable amount of time to solve the problem. If a KR system were able to pick them out as a possible match for a kinship relation, it could be of great help. Upon request, a picture could be taken of a refugee and put in a database. This database could then check for kinship relations with other refugees in the same database. If a missing person is registered in a camp, a kinship relation could be detected instantly this way. They would be reunited much faster and more efficiently. Issues such as communication and limited manpower could also be reduced with the discussed automation.

Another example of the usage of KR is focused on safety measures. Imagine a situation where a terrorist is not in the system. No information can be found on them, only an image of their face is accessible. KR systems could try to match possible family members that are in the system of known suspects. This could lead to finding the terrorist sooner or to finding their accomplices. Like this, there are more situations where automated KR would be really helpful. With the reality of these problems, KR can bring families together and provide more safety.

The main contribution of this paper is to make a first step towards understanding automated KR and the importance of facial features in it. In the field of KR, there is a lot of room for improvement, especially on the importance of facial features.

In our research, we tackled whether kinship is recognizable by using a set of extracted facial features with the use of machine learning. Specifically, we focus on what specific set of features is important for automated KR and if this set of features complies with the set of features important in human KR. First presented in this paper is a literature discussion on human as well as automated KR. We then discuss the data in Section II. In Section III, an overview of the used models is presented. Next, we discuss the results of different experiments in Section IV. Lastly, a discussion and conclusion of the presented experiments are given in Sections V and VI.

B. Related work

1) Kinship Versus Look-alike: In various researches, which we will discuss later, it has been shown that automated KR is possible to a certain degree. The main question we are left with is how we would separate the classification of two family members from two people that look alike [7], [8]. On one side, two people could be unrelated but their faces as a whole could look alike. If their facial features would be extracted and compared, the features would probably not have high similarity [9]. They have lower inter-class variations. Inter-personal variations refer to the differences in race or genetics. This includes variations such as eye color and the shape of a nose, features that are not possible to be (easily) changed. On the other side, intra-personal variations refer to variations in features that are easily changed, such as hair, facial accessories, cosmetics, pose and illumination [8]. Their face looks similar due to these intra-personal variations. With this information, we expect each feature separately to not show significant similarities. For example, having a close look at the shape of their nose, it is considerably different from the shape of the other person's nose. The intra-class similarity is higher and the inter-class variation is lower for the two individuals of this example [8].

On the other side, there are two people that are related, a daughter and father for example. They do not necessarily look alike since they differ in age and sex. They are in general not seen as lookalikes. However, when looking at their facial features, most of the time you would see that there is a high similarity for some features, whereas other features would not be similar at all [10]. An example of this is a father and daughter who have a nose and mouth that are very similar. However, their faces as a whole do not look alike, because the rest of their facial features are not similar at all. This would be due to the heredity in kinship, as not all traits inherited from a parent to a child are reflected in the child's appearance.

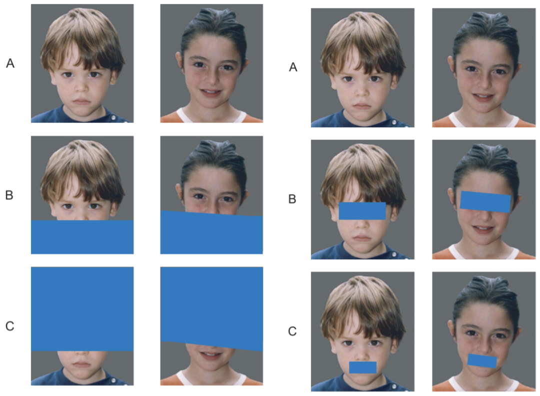
2) Human Kinship Recognition: Studies on human KR contribute to our search for the set of important features in automated KR. Several studies [11]–[13] have been conducted on human KR, which showed that kinship is indeed recognizable by humans. Robinson et al. [14] used the Families-In-the-Wild (FIW) data set for their human performance measurement. This data set contains images of people's faces that are extracted from family pictures. In total, the data set consists of $656,954$ pairs of images that show a kinship relation. Robinson et al. state that humans scored an overall average of 56.6% accuracy. In this experiment, two pictures were shown and a binary classification was performed between related by kinship and unrelated. Other research on KR [11], [12], [15], [16] shows similar results. The results from Lu et al. [14] are shown in Table I. They performed two different experiments to test human KR. The test group is split up into group A and B. Group A was only shown a cropped face region, whereas group B was shown the whole original color images. Group A intends to test kinship verification purely based on face, while group B intends to test kinship verification based on multiple cues including face, hair, skin color, and background [14]. Their results show that the average accuracy of human KR is higher when face, hair color, and background are taken into account compared to when the focus is purely on the face.

TABLE I. RESULTS OF THE TWO EXPERIMENTS ON HUMAN KINSHIP RECOGNITION FOCUSED ON FOUR KINSHIP TYPES BY LU ET AL. [14]. THE NUMBERS REPRESENT THE ACCURACY AND THE DIFFERENT COLUMNS REPRESENT THE KINSHIP RELATIONS FATHER AND SON (F-S), FATHER AND DAUGHTER (F-D), MOTHER AND SON (M-S), AND MOTHER AND DAUGHTER (M-D).

Method	F-S	F-D	M-S	M-D	Mean
HumanA	61.00	58.00	66.00	70.00	63.75
HumanB	67.00	65.00	75.00	77.00	71.00

We take a look at the Feature Importance (FI) in some of these studies on human kinship detection. The reason behind this specific set of features for human KR might be of help in automated KR. One of the studies is by Martello and Maloney [11], [12], who raised the question which parts of a face are most important for human KR. In [11], a study is conducted in which humans were tested on their KR skills based on three separate conditions: (1) the right hemi-face masked, (2) the left hemi-face masked, and (3) the face fully visible. Most interestingly, the results showed that there is no significant difference in results for recognizing kinship by humans when the left or right part of the face is covered. On the contrary, a similar study [12] showed that the covering of the top or bottom part of a face does give a significant difference. The effect on kin recognition performance of masks that covered the upper half or the lower half of the face (experiment 1) and the eye region or the mouth region (experiment 2) were measured. An example of the covering up of facial parts for experiments 1 and 2 can be seen in Figures 1a and 1b below.

In these experiments, it was found that masking the eye region led to a 20% reduction in performance whereas masking the mouth region did not yield a significant difference in performance. This leads us to consider the theory that the performance in KR is dependent on only the upper half of a person's face. Curious is to see how this theory could be used in automated KR. Another discovery is that the eye region contains only slightly more information about kinship than the upper half of the face outside of the eye region. Moreover, the theory is discussed that splitting up face images in different patches can improve the ability of humans to recognize kinship. This would be caused by the mouth area



(a) Experiment 1: Masking the bottom and top half of the face

(b) Experiment 2: Masking the eye area and lip area of the face

Figure 1. Illustration of the masking of faces in [12].

(i.e., the bottom half of the face) containing lesser kinship features as it is subjected to considerable changes during development [17]. The lower face does not reach its final form until early adulthood [18]. Consequently, this area has fewer stable cues to relatedness. Another view discussed [17] is that environmental effects have little influence on the detection of kinship using facial similarities. This indicates that genetically irrelevant facial information is ignored when human KR is performed.

Overall, the theory that we researched is based on the change in performance when using a specific set of facial features compared to facial features from the whole face. The theory implies that there is a mere necessity of these facial features for KR. These are the features that are located in the upper half of the face, which could lead to only requiring specific parts of faces to identify kinship relations. This could lead to more accessible data since only parts of faces are also sufficient for extracting the important features. Moreover, it could decrease the computational cost of KR models.

3) Automated Kinship Recognition: For *automated* KR, several approaches have been proposed. Most approaches are not only focused on machine learning models, but also on feature selection. Feature-based methods aim to preserve facial, genetically determined characteristics in the feature descriptors used for the model. These methods identify local facial features such as inconsistencies in an individual's eyes, mouth, nose and skin from the individual's image. Feature-based methods can decrease computational costs and improve model performance. Most of the proposed models and algorithms were only trained on small data sets.

These are data sets like KinFaceW [14], [19] where only four types of kinship relations were given on a handmade data set of around 150 images [20]. Another data set in the field of KR is TSKinFace (Tri-Subjects Kinship Face Database). This data set has been used in some studies [21], [22], but also proved to be too small. These data sets demonstrated to be insufficient for the task at hand. Most of the proposed classifiers are lower-level models and algorithms which use handcrafted feature extraction (features using information presented in the image itself), Support Vector Machines or K-Nearest Neighbor classifier.

Since 2016, a more extensive data set has been constructed in [13]: Families-in-the-Wild (FIW). This data set has been produced to verify kinship and classify relations [23]. The creators of this data set specify promising results in detecting kinship. Robinson et al. [14] state the best results were obtained when using the SphereFace model with an average accuracy of 69.18% and standard deviation of 3.68. All models performed well compared to previous work, although much improvement could still be made.

After publishing the FIW data set, more research in the field of KR models was done. Many models in KR include the use of FaceNet or other small feature selections for their models' input [24]. FaceNet is a neural network that extracts features of an image. The model provides a mapping from a picture of a face to the Euclidean space. The distances in this space correlate to the amplitude of face resemblance [25]. It produces an output vector to be used as input for a classification model. FaceNet creates embeddings by learning the mapping from images. A disadvantage of using FaceNet is that especially when looking at FI, information gets lost due to lack of feature interpretation [26]. FaceNet could help improve KR models, although we are interested in the similarities between faces by using facial features instead of the faces as a whole. Hence, we use a different approach than FaceNet.

Fang et al. [27] proposed different feature extractions. They performed classification on a pair of images based on the difference between feature vectors of the pair. These pairs are a potential parent and child pair. The top selected features showed to be right eye RGB color, skin gray value, left eye RGB color, nose-to-mouth vertical distance, eye-to-nose horizontal distance and left eye gray value. The results show a high importance for eye related features. 10 out of the 14 top features include the eye area. This study does include specific facial features like eye color, still it only included 22 low-level features. It is indeed shown that most of the selected features are in the upper face area, which complies with the hypothesis.

The top selected features are right eye RGB color, skin gray value, left eye RGB color, nose-to-mouth vertical distance, eye-to-nose horizontal distance and left eye gray value. The results show a high importance for eye related features. 10 out of the 14 top features include the eye area. While this study does include specific facial features like eye color, it only included 22 low-level features. It is indeed shown that most of the selected features are in the upper face area, which complies with the insight.

Most studies on the subject focus on either the overall similarities between faces, or on pre-determined facial feature sets. These studies treat KR tasks similar to the task of a standard facial recognition. Guo et al. [28] argue that kinship classification should be treated differently, since trait similarities are measured across age and gender. Additionally, kinship has a combination of traits and familial traits, which are special for each family pair.

Models proposed by researchers in this field are based on an input of just the images with little to no alterations. Although, some research focus on specific facial features by using for example a weighted graph embedding-based metric learning framework [31] or by using sparsity to model the genetic visible features of a face [32].

Another group of researchers thought of combining the StyleGAN2 algorithm with KR [33]. In the task at hand, there is a restriction that family members should be recognized on the basis of physical facial features. However, several mentioned attempts neglect this constraint and do not employ any facial landmark before using a classification model. For this reason, Nguyen et al. [33] experimented with KR models using StyleGAN2 as an encoder to incorporate a facial landmark map. This method resulted in an average accuracy of 0.548 for recognizing kinship. Against expectations, no improvement was shown in the results from using StyleGAN2 in this manner, which is presumed to be due to the lack of a proper classification and thus it is argued to need more investigation. An algorithm proposed by Guo et al. [28] use familial traits extraction and kinship measurement based on a stochastic combination of the familial traits. The authors use a similarity score based on a Bayes decision for each pair of facial parts. However, facial features used by the algorithm are limited to the eyes, nose and mouth and, in line with the observations by Guo et al. [28], more parts of the face could be explored. Existing data sets use faces from the same family picture, so models learn about the background similarity. This causes the models to get a higher performance, but when tested on real life pictures, not taken from a family picture, the performance could be lower. When using pre-determined features, this does not present a problem.

II. Data

We used the Families in The Wild (FIW) data set. FIW is made up of 11,932 natural family photos of 1,000 families. Other data sets contain less images. KinFaceI consists of 1000 pairs of pictures (so even less unique pictures) [29] and TSKinFace consists of 787 pictures [30]. This makes the FIW data set by far the biggest data set being used for KR.

The data contains images of people's faces that are extracted from family pictures, hence the images vary in quality. All images of the persons are of the same size (108x124 pixels). Some pictures are zoomed in on more than others, which causes this quality difference. In some images, the face and its facial features are clearly shown, but other images are very blurry.

The data is split up into training and test data using hold-out cross-validation. The data is split up in a 70/30 split, respectively. The training set consists of information on families, persons and relations between persons including images of the persons. The data is distributed as follows. An average of about 12 images per family, each with at least 3 and as many as 38 members. Each family is assigned a unique id, each person is assigned an id and each image collected is assigned a unique id. The data set includes good-quality images of a person's face, but also blurry images of faces, as shown in Figures 2a and 2b, respectively.

(a) Image from data set

(b) Blurry image from data set

Figure 2. Example data from the Families-in-the-Wild data set

A file containing all matches in the training data set is available. However, this does not include data on combinations of persons that do not have a familial relationship. So, these pairs have been constructed by taking random pairs of images from the set of training images of the FIW data set. This is excluding existing related pairs and each pair is unique. This resulted in 205,285 related and 205,285 unrelated pairs of images. The kinship relations are labeled as related. Of all related data points, 21% of the data points are zeroth generation (siblings), 75% are first generation (parents and children) and 4% are second generation (grandparents and grandchildren).

Binary classification is used for predicting relatedness, so the data set is balanced accordingly. Since the focus is on the importance of each of the facial features in recognizing kinship, we have a split in the data between related and unrelated. The distinction between the types of kinship relations is not made. For now, the types are not taken into the classification process, considering the aim is to have a general interpretation of the important facial features. However, the distribution between the types of pairs can help to understand the possible patterns found in feature importance.

StyleGAN2 metric: linear separability

This research is focused on FI in KR. To be able to understand the FI of a model, the features extracted from a model should be interpretable. To collect a bunch of features and to avoid having to do manual annotation, we decide to use a feature description method from the StyleGAN2 model. With this, it can be easily deducted which of the features of a face are seen as most important by a model for detecting kinship. The pictures in the data are of size 108x124, while the StyleGAN2 description method expects pictures of size 256x256 as input. Interpolation of the pictures in the data is used to overcome this problem. The StyleGAN2 model contains a certain metric called linear separability. StyleGAN2's linear separability metric can be used to steer a generated picture in a certain direction by specifying 40 facial features which are shown in Table IV. For example, the models can be used to make the generated face have blond hair and high cheekbones. What we are most interested in for this research are the pre-trained models used in StyleGAN2 which produce probabilities of the 40 features to be true for an image of a person.

TABLE II. FACIAL FEATURES OF LINEAR SEPARABILITY METRIC

1) 5-o-clock-shadow,
2) arched eyebrows,
3) attractive,
4) bags under eyes,
5) bald,
6) bangs,
7) big lips,
8) big nose,
9) black hair,
10) blond hair,
11) blurry,
12) brown hair,
13) bushy eyebrows,
14) chubby,
15) double chin,
16) eyeglasses,
17) goatee,
18) gray hair,
19) heavy make up,
20) high cheekbones,
21) male,
22) mouth slightly open,
23) mustache,
24) narrow eyes,
25) no beard,
26) oval face,
27) pale skin,
28) pointy nose,
29) receding hairline,
30) rosy cheeks,
31) sideburns,
32) smiling,
33) straight hair,
34) wavy hair,
35) wearing earrings,
36) wearing hat,
37) wearing lipstick,
38) wearing necklace,
39) wearing necktie,
40) young.

The metric was trained using the CelebA Data set (CelebFaces Attributes Data set). This is a face attributes data set with 202,599 celebrity images, each with five landmark locations and 40 attribute annotations. StyleGAN2's linear separability metric is meant to be used for the StyleGAN2 model and its corresponding data. We are interested in using the metric on the data from FIW. The information gathered from the linear separability metric (the facial features) is used as a starting point for the kinship classification models. Transfer learning does not only save time, but it also has the possibility of making a learning process more efficient [34].

Consequently, some adjustments to the data were necessary to apply the metric. This resulted in an output of 40 features for all images in the data set, which then could be used to train the chosen automated KR models. As data points for the models, we chose a list of length 40 and a list of length 80, composed of the metric values for the features per two pictures. Two input types were experimented with: (1) a list of 80 features, consisting of 40 features per image, and (2) a list of 40 features, taking the absolute difference of the feature values between the images per feature.

III. Model description

We implemented and tested several models to see how well the models work on our data and to find a recurring pattern in FI. For all models, the FI is investigated. The results of this are then used to understand whether the theory of human KR will hold for automated KR as well. Various machine-learning models were selected for this task. For each model, the accomplished accuracy is obtained by K-fold cross-validation. The number of folds is set to 10 and the data is shuffled before splitting into batches.

Machine learning methods

Using StyleGAN2's linear separability metric on our data results in an output of 40 features for all images in the data set, which then are used to train the models. As data points for the models, we chose a list of either length 40 or 80, composed of the metric values for the features per two pictures. The five models we decided to experiment with are decision tree, random forest, Gaussian naive Bayes, linear support vector machine and logistic regression.

Decision Tree: First, we have the decision tree algorithm with a maximum depth set to 10, where we obtain the FI by using the Gini importance. The Gini importance is calculated as shown in Equation (1) with ni_j the importance of node j, w_j the weighted number of samples reaching node j, C_j the impurity value of node j and $left(j)$ and $right(j)$ the child nodes from left and right split respectively on node j.

$$ni_j = w_j C_j = w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \qquad (1)$$

The Gini importance value ni_j for feature i is then used for the feature importance of feature i with Equation (2) where fi_i is the importance of feature i and ni_j the importance of node j. These values were then normalized to a value between 0 and 1 by dividing by the sum of all feature importance values [35].

$$fi_i = \frac{\sum_{\text{j: node j splits on feature i}} ni_j}{\sum_{\text{k}\in \text{ all nodes}} ni_k} \qquad (2)$$

Decision trees are easily interpretable. Because of the use of decision-making logic, the information on the model's features is easily extracted in a comprehensible form [36]. Decision trees have a built-in feature selection, which is beneficial for our research [37]. However, overfitting is common when using decision trees. This is due to the trees being too complex.

Random Forest: Second, we have the random forest consisting of 100 trees, where the FI is obtained by using the impurity importance. The feature importance is computed as an average over all trees. The splitting rules of a random forest scale down the impurity presented by a split. When a split shows a considerable decrease in impurity, the split is seen as important. This theory results in the impurity importance calculation for a variable in the random forest as shown in Equation (3) where $RFfi_i$ is the importance of feature i in the random forest, $normfi_{ij}$ the feature importance for feature i in tree j normalized, T_{all} the set of all trees, and T the number of trees in the random forest [35], [38], [39].

$$RFfi_i = \frac{\sum_{j \in T_{all}} normfi_{ij}}{T} \qquad (3)$$

Where decision trees are susceptible to overfitting, specifically when a tree is notably deep, the random forest algorithm reduces this possibility of overfitting. This is due to random forest algorithms constructing multiple decision trees where more combinations of conditions are represented [40].

Gaussian Naive Bayes: Then, we have the Gaussian Naive Bayes, which obtains FI by using the permutation importance. The permutation importance is calculated by taking the difference between the prediction error of the baseline metric and the prediction error of the permuted feature metric as shown in Equation (4). The permutation importance is obtained as follows. First, the model m is scored s on data D. Then permutation variable importance of feature j is calculated. For each feature, the feature column is randomly shuffled to create an adjusted version of the data $\hat{D}_{k,j}$. The model is then again scored on the data, although now with the feature f replaced by the adjusted version. This results in the score $s_{k,j}$, after which the importance i_j for feature f_j can be calculated with Equation (4) [41].

$$i_j = s - \frac{1}{K}\Sigma_{k=1}^{K} s_{k,j} \tag{4}$$

This algorithm generally works very fast and can easily predict the class of a test data set. It is not sensitive to irrelevant features [42]. The naive Bayes algorithm does perform the best overall when there is independence between the features, while some of our features are dependent. It assumes that all the features are independent [43]. However, even without independence between the features, the naive Bayes algorithm generally performs well.

Linear Support Vector Machine: Next is the linear Support Vector Machine (SVM), where the weights of the model are used to determine FI. These weights are used as vector coordinates. The vector coordinates are orthogonal to the hyperplane represented by the weights. The directions of the vectors represent the class prediction. The difference in the size of the weights is used to determine the feature importance [44]. SVMs have a low risk of overfitting [45], outliers have less influence in the algorithm and the SVM algorithm is relatively memory efficient [46]. Nonetheless, understanding the final SVM model and interpreting the feature importance is difficult. Additionally, SVMs are usually not very suitable for large samples of data, although LSCVs handle this better [47].

Logistic Regression: Lastly, we have logistic regression, where the FI is determined by using the coefficients of the decision function. To get a feeling for the "influence" of a given parameter in a linear classification model, the magnitude of the coefficient for each feature times the standard deviation of the corresponding parameter in the data is considered. The positive coefficients correspond to outcome 1 (related) and the negative coefficients correspond to outcome 0 (unrelated). This means that a higher positive value of the corresponding feature pushes the classification more towards the negative class [48]. Logistic regression is easy to implement and interpret and very efficient to train. Therefore, it does not require high computation power [49]. The algorithm does make the assumption of linearity between the log odds and the independent variables [50].

IV. Results

Three different approaches have been researched, the original StyleGAN2 description method, the pre-selected features method and the bottom and top masked method. The results of these approaches are discussed and an overview of the results is provided.

A. Original StyleGAN2 descriptor experiment

The initial approach is taking the results of the StyleGAN2 model and using them as input for the different algorithms. Over all images, we calculated the probabilities of the image complying with the given 40 features. Extracting 40 features per picture resulted in 80 different values since we were working with two images per data point. The FI was determined per model. For the 80 feature input, we took the sum of each feature per picture. An overview of all the results from the StyleGAN2 descriptor experiment can be found in Table VI and Table VII.

Decision Tree: The accuracy of the decision tree with 40 features as input has a mean of 0.61 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.66 with a standard deviation of 0.005. The model is more leaning towards giving a positive (related) classification. The feature importance for the decision tree model is shown in Figure 3. For the decision tree model with input of 40 features, *arched eyebrows*, *no beard* and *heavy makeup* are the most important features. For the input of 80 features, the top most important features include *young*, *no beard* and *wearing necklace*.

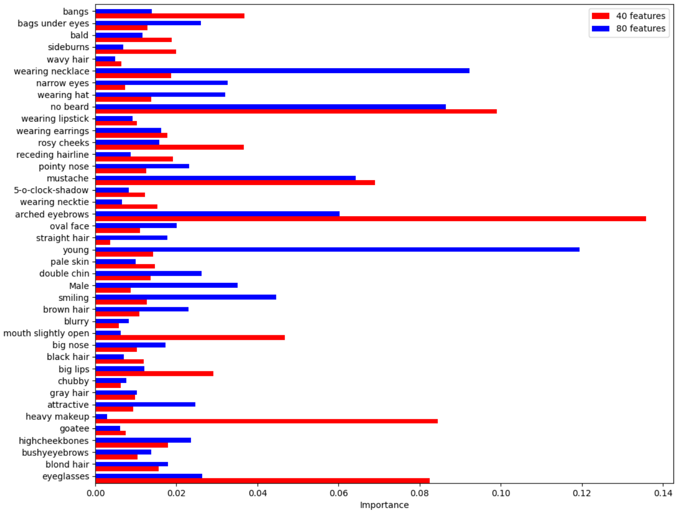


Figure 3. Barplots of feature importance for the decision tree model.

Random Forest: The accuracy of the random forest with 40 features as input has a mean of 0.74 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.80 with a standard deviation of 0.004. The model does not have a clear preference for either a positive or negative classification. With the model giving 51.39% and 50.63% positive classifications for 40 and 80 features respectively, the even distribution of the data in half-positive and half-negative data points is represented well with a slight deviation towards positive classifications. The feature importance for the random forest model is shown in Figure 4. For the random forest model with input of 40 features, *arched eyebrows*, *mustache* and *heavy make up* are the most important features. For the input of 80 features, the top most important features include *young*, *no beard* and *mustache*.

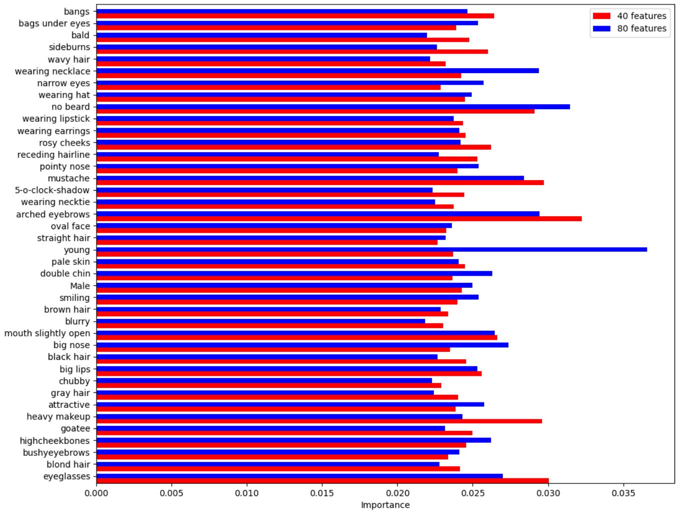


Figure 4. Barplots of feature importance for the random forest model.

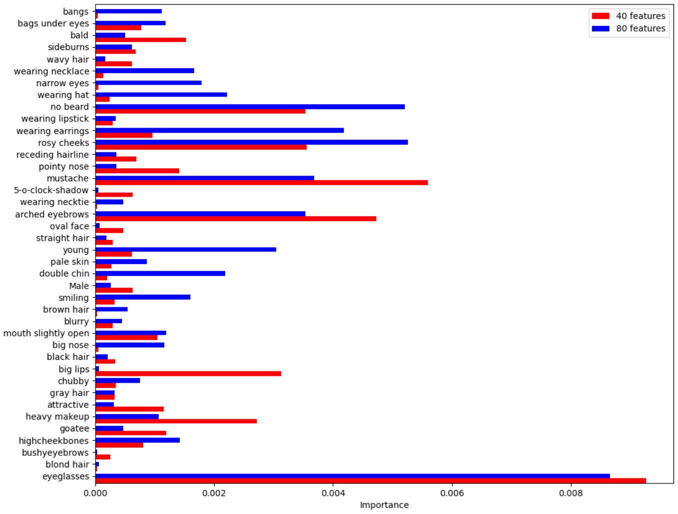


Figure 5. Barplots of feature importance for the naive Bayes model.

Gaussian Naive Bayes: The accuracy of the Gaussian naive Bayes with 40 features as input has a mean of 0.60 with a standard deviation of 0.004. The 80 features input gives a mean accuracy of 0.59 with a standard deviation of 0.005. The model has a preference for positive classification. With the model giving 59.56% and 64.21% positive classifications for 40 and 80 features respectively, most errors are false positives. The feature importance for the Gaussian naive Bayes model is shown in Figure 5. For the Gaussian naive Bayes model with input of 40 features, *eyeglasses*, *mustache* and *arched eyebrows* are the most important features. For the input of 80 features, the top most important features include *eyeglasses*, *rosy cheeks* and *no beard*.

Linear Support Vector Machine: The accuracy of the linear SVM with 40 features as input has a mean of 0.59 with a standard deviation of 0.004. The 80 features input gives a mean accuracy of 0.63 with a standard deviation of 0.005. The model does not have a clear preference for a positive or negative classification. With the model giving 47.08% and 52.92% positive classifications for 40 and 80 features respectively, we see a slight effect of the different input values. The 40 values input gives the model a bit more lenience towards negative classification and the 80 values input gives the model slightly more lenience towards positive classification. The feature importance for the LSVM model is shown in Figure 6. For the LSVM model with input of 40 features, *arched eyebrows*, *no beard* and *heavy make up* are the most important features. *no beard* and *arched eyebrows* are also among the most important features for the input of 80 features. Here the top most important features include *arched eyebrows*, *narrow eyes* and *no beard*.

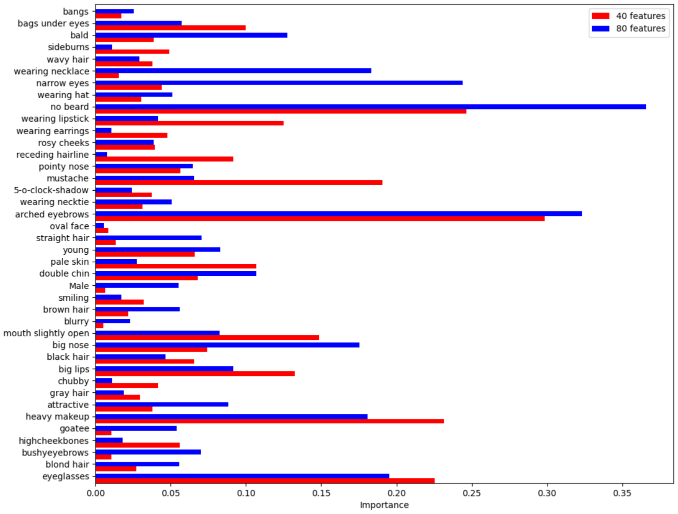


Figure 6. Barplots of feature importance for the LSVM model.

Logistic Regression: The accuracy of the logistic regression with 40 features as input has mean 0.60 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.63 with a standard deviation of 0.005. The model does not have a clear preference for a positive or negative classification. The model gives 50.40% and 51.61% positive classifications for 40 and 80 features respectively, which shows the balance of the data with a slight deviation towards positive classification. The feature importance for the Logistic Regression model is shown in Figure 7. For the logistic regression model with input of 40 features, *arched eyebrows*, *no beard* and *eyeglasses* are the most important features. These are also among the important features for the input of 80 features. Here the top most important features include *no beard*, *arched eyebrows* and *pale skin*.

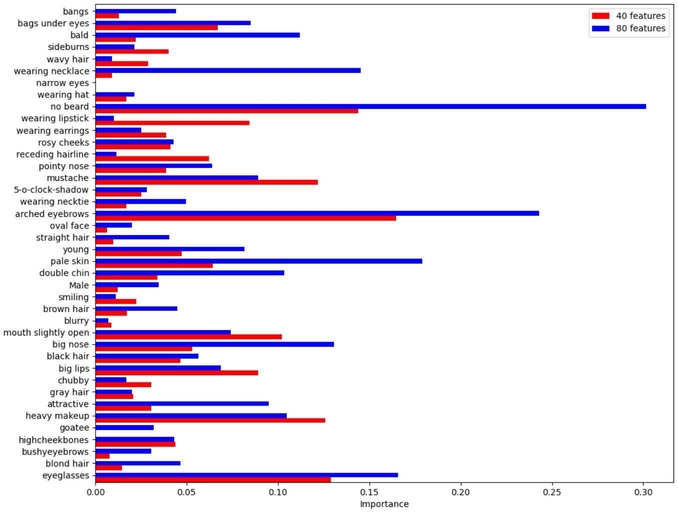


Figure 7. Barplots of feature importance for the logistic regression model.

B. Selected StyleGAN2 descriptor experiment

The second approach is based a certain selection of features. Two different selections were experimented on. The first is focused on the human KR theory. A pre-selection of features is applied to the selected models. The selection of features is focused on the top half of a face. This selection is shown below in Table III.

TABLE III. SELECTED SET OF TOP HALF FACIAL FEATURES OF LINEAR SEPARABILITY METRIC

1) Wavy hair,
2) 5-o-clock-shadow,
3) arched eyebrows,
4) bags under eyes,
5) bald,
6) bangs,
7) black hair,
8) blond hair,
9) brown hair,
10) bushy eyebrows,
11) eyeglasses,
12) gray hair,
13) high cheekbones,
14) narrow eyes,
15) receding hairline,
16) sideburns,
17) straight hair,
18) wearing earrings,
19) wearing hat.

This approach, despite the supporting theory, did not give better results than using all the features. The expectation was stability in the accuracy scores by only focusing on the allegedly most important features. However, most models were less successful and only some of the models continued to perform roughly the same. The accuracy scores for the models using the 19 selected features are given in Table V. For this experiment, the differences were taken between the features, so the results should be compared to the results of the method where the input was 40 features as shown in Table V.

The second approach is based on the selection of features found to be most important according to the original StyleGAN2 approach. The top most important features for this approach can be found in Table IV. When using only these features as input, the results are as shown in Table V.

TABLE IV. SET OF MOST IMPORTANT FOUND FACIAL FEATURES OF LINEAR SEPARABILITY METRIC

1) Arched eyebrows,
2) eyeglasses,
3) heavy makeup,
4) mustache
5) narrow eyes,
6) no beard,
7) young.

TABLE V. OVERVIEW OF RESULTS, PRESENTED AS ACCURACY, FOR THE TWO SETS OF SELECTED FEATURES COMPARED TO THE RESULTS OF ALL FEATURES

	Selection Top	**Selection Important**	**All 40**
Decision Tree	0.61	0.61	0.59
Gaussian Naive Bayes	0.57	0.59	0.61
Support Vector Machine	0.57	0.57	0.60
Logistic Regression	0.57	0.57	0.60
Random Forest	0.71	0.62	0.74

C. Masked StyleGAN2 descriptor experiment

To support the theory we found, all of StyleGAN2's linear separability features were taken of not the original image, but over an image with the bottom part of the face masked black like shown in Figure 8. The same was done with the top part of the face masked black, comparable to the experiments performed by Martello et al. [11], [12]. All the models are exactly the same as for the original StyleGAN2 description method. Only the input changed.

(a)

(b)

Figure 8. Example data from the Families in the Wild data set with bottom masked (a) and top masked (b)

Bottom half masked: This experiment was done with all models previously used in the original StyleGAN2 descriptor experiment. The accuracy and FI were obtained for the decision tree, random forest, Gaussian naive Bayes, LSVM and logistic regression models. An overview of the accuracy and important features for all the models from the bottom masked StyleGAN2 descriptor experiment can be found in Table VI and Table VII. Again, the results show that the 80 value input gives an overall better performance than the 40 value input and the best-performing model is the random forest for both inputs. Some of the most important features for the bottom masked approach are related to the nose (*pointy nose* and *big nose*) and the hair (*grey hair*, *blond hair* and *waivy hair*).

Top half masked: This experiment was done with all models previously used in the original StyleGAN2 descriptor experiment. The accuracy and FI were obtained for the decision tree, random forest, Gaussian naive Bayes, SVM and logistic regression models. An overview of the accuracy and important features for all the models from the bottom masked StyleGAN2 descriptor experiment can be found in Table VI and Table VII. Again, the results show the 80 value input gives overall better performance than the 40 value input and the best performing model is the random forest for both inputs.

TABLE VI. ACCURACY FOR THE 40 AND 80 VALUE INPUT PER EXPERIMENT: COMPLETE, BOTTOM MASKED AND TOP MASKED

	40 Compl.	**40 Bottom**	**40 Top**	**80 Compl.**	**80 Bottom**	**80 Top**
Decision Tree	0.61 ± 0.003	0.57 ± 0.004	0.57 ± 0.003	0.66 ± 0.005	0.64 ± 0.004	0.65 ± 0.003
Random Forest	0.74 ± 0.003	0.62 ± 0.003	0.63 ± 0.002	**0.83 ± 0.004**	0.81 ± 0.001	0.82 ± 0.001
Gaussian Naive Bayes	0.60 ± 0.004	0.53 ± 0.003	0.55 ± 0.002	0.59 ± 0.005	0.55 ± 0.003	0.57 ± 0.002
Support Vector Machine	0.59 ± 0.004	0.55 ± 0.002	0.57 ± 0.002	0.63 ± 0.005	0.60 ± 0.002	0.61 ± 0.002
Logistic Regression	0.60 ± 0.003	0.55 ± 0.003	0.57 ± 0.002	0.63 ± 0.005	0.59 ± 0.003	0.61 ± 0.002

TABLE VII. MOST IMPORTANT FEATURES PER EXPERIMENT

	Complete	**Bottom Masked**	**Top Masked**
Decision Tree	young, no beard, arched eyebrows, eyeglasses	attractive, blond hair, pointy nose, grey hair	young, no beard, arched eyebrows, eyeglasses
Gaussian Naive Bayes	eyeglasses, no beard, young, arched eyebrows	wavy hair, blond hair, pale skin, heavy makeup	eyeglasses, no beard, young, arched eyebrows
Support Vector Machine	young, no beard, pointy nose, arched eyebrows	grey hair, pale skin, wavy hair, big nose	young, no beard, pointy nose, arched eyebrows
Logistic Regression	blurry, no beard, wearing necklace, pointy nose	wavy hair, young, grey hair, big nose	blurry, no beard, wearing necklace, pointy nose
Random Forest	young, no beard, mustache, arched eyebrows	pointy nose, grey hair, smiling, attractive	young, no beard, mustache, arched eyebrows

V. Discussion

Multiple models have been tested on FI. Some approaches were based on the human KR experiments from [11], [12]. These experiments showed a certain area of the face to contain the important facial traits needed for KR. We researched the set of features that is most important for automated KR. Pretrained metrics from the StyleGAN2 model that are meant to be used for synthesizing artificial examples of faces were used. The pre-trained models give 40 values for specific facial features. These 40 values can also be taken from pictures using the pre-trained models. These values were used as input for our machine learning models: decision tree, random forest, Gaussian naive Bayes, support vector machine and logistic regression. These models were trained and evaluated to show which of the features were seen as most important by the models. More experiments were conducted with the top and bottom parts of a face masked black to also test the theory of human KR.

Major findings: Interesting results were found when comparing the different models using the original StyleGAN2 description method. Four out of five models had a higher accuracy score when all features for both pictures were kept separate. The models are able to learn about combinations of different features between the two pictures, which has a positive influence on the accuracy score of the models.

The best-performing model seems to be the random forest. Since this model has a very high accuracy compared to the other models, we are specifically interested in its corresponding FI scores. Accordingly, we mainly focus on the results of the random forest model. This model gives high importance values to the features *young*, *no beard*, *mustache* and *arched eyebrows*. It is also noticeable that in two of the five models, the feature *young* is found to be very important and in the other three models, the FI increases when using 80 features instead of 40 features as input. On top of that, in all models, the features *arched eyebrows* and *no beard* are in the top four of the most important features for the model. There is a clear pattern in the importance of facial hair. Beards, mustaches and arched eyebrows are found to be important features for most of the models. Another pattern is the age difference. This gives us reason to believe that the combination of facial hair and the age of a person is strongly correlated to the classification. While the correlation scores do not show a correlation between the two features, the combination of the features does matter when comparing two pictures. A reason for these features to be found important is that most of the kinship relations (75%) in the data set are zero-generation and first-generation relations. Young people are not able to grow facial hair, if they have the genes, it comes with age. This would explain why both facial hair and age are found to be more important.

The set of features that were found to be the most important in our research does not comply with the selection of features proposed by Fang et al. [27]. The set of features used in their research is different, although it is clear that the eye area was found to be the most important by them. Contrasting, the set of important features we found is not particularly focused on the eye area.

For the masked experiment, all five models had a higher accuracy score when all features for both pictures were kept separate. When looking at the bottom masked method results, a clear decrease in the performance is found compared to the original StyleGAN2 description method. Remarkable is that the feature *young* and the features on facial hair are not found in the top features of almost all models. The original StyleGAN2 approach showed these features to be very important. This leads to the believe that the bottom part of a face is essential for extracting the feature *age*. This would also explain why the feature *grey hair* is found to be important in three out of five models. Grey hair is usually a sign of a higher age. When looking at the top masked method results, a decrease in the accuracy is found, although this decrease is not as excessive as with the bottom masked method. Above is mentioned that the feature young is likely to be extracted from mostly the bottom of a face. However, this is not shown in the results of the top masked method. It is curious that the feature *young* is still not found to be one of the most important. Like the original approach, the top masked method shows the feature *arched eyebrows* to be important. Although a pattern is difficult to find in the top masked method results.

For the bottom masked approaches the difference with the original approach is clear. Where humans showed equal or even better performance when masking the top half of a picture, the algorithms showed the opposite effect.

The set of features does not comply with the set that we expected it to comply with. The gathered results do not give any information that would validate the hypothesis that the most important features would be in the upper half of the face, specifically the eye region. On the contrary, the results are more lenient towards age and facial hair traits to be of great importance. As for the approach with the pre-selected StyleGAN2 linear separability features, the results showed us no improvement when focusing on solely the upper half of the face. When considering the results of the pre-selected StyleGAN2 and the masked StyleGAN2 descriptor experiments, rejecting the hypothesis is even more reasonable.

Limitations: The data set might not be very compatible with the StyleGAN2 metrics, which is an uncertainty. However, as of now, there are no other data sets that contain enough images which are of adequate quality. So we have to accept this limitation for now. An issue was also encountered when using the linear separability metric for a different purpose than StyleGAN2. The results for the top masked method showed one very noteworthy important feature, namely the *arched eyebrows* feature. This feature should be focused on the top part of a face. However, it is found to be important when the top part of a face is masked. More features which show unusual behavior are *smiling* and *pointy nose*, since these are found to be important when masking the bottom half of the face. This is one of the problems that is encountered when combining StyleGAN2 metrics with other models. The models that are trained for the linear separability metric behave differently than intuitively expected. Using the metric in tasks for which it is not initially intended can cause limitations to the models.

Unexpected findings: A surprising matter is the difference in performance between the top masked and bottom masked StyleGAN2 description method. Masking the bottom half of the face decreased the performance. As masking the top half of the face decreased the performance as well, it still performed better than the bottom masked method. This is against expectations and raises the question of whether the bottom part of a face contains more information than the top part of a face does for KR.

VI. Conclusion

We researched the set of features that is most important for automated KR. For this, multiple models have been tested on FI. The results showed that the most important facial features from the selection of 40 features are mostly focused on the facial hair traits and age-related features.

One of the issues we ran into is on transfer learning. The question rises whether StyleGAN2 is compatible enough for transfer learning when combined with our data set. It could be more effective to write a new metric that focuses on more solid facial features. Despite that, the StyleGAN2 metrics are the most elaborate method for finding pre-determined facial features. Other models do not include as many facial features or need manual annotation. It would be contributory to find a way to annotate all parts of the face for many more features to train the models on.

In conclusion, this paper is an important first step towards understanding automated KR, but there are many challenges to be faced before it can be used in real-world applications. As it is now, a large set of clear pictures of complete faces are needed for a model to perform decently. Learning more about the most important parts of our face for automated KR is the next step to take to improve the field of KR.

VII. Acknowledgment

We thank Rob van der Mei and Sandjai Bhulai for their useful comments on drafts of this paper.

References

[1] B. E. van Leeuwen, A. Gansekoele, J. Pries, E. van de Bijl, and J. Klein, “Explainable Kinship: The Importance of Facial Features in Kinship Recognition”, Iaria Congress 2022 Proceedings, pp. 54-60, 2022

[2] J. Brownlee, “Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python”, https://books.google.nl/books?id=DOamDwAAQBAJ, Machine Learning Mastery, 2019.

[3] R. Szeliski, “Computer Vision: Algorithms and Applications”, https://books.google.nl/books?id=bXzAlkODwa8C, Springer London, 2010.

[4] S. A. Papert, “The Summer Vision Project”, http://hdl.handle.net/1721.1/6125, 1966.

[5] “A brief history of facial recognition - NEC New Zealand”, https://www.nec.co.nz/market-leadership/publications-media/a-brief-history-of-facial-recognition/, May 2020, (Accessed on 21/10/2022).

[6] E. Seselja, “How the Red Cross and a radio reconnected a family torn apart by conflict - ABC news”, https://www.abc.net.au/news/2021-08-29/red-cross-reconnect-family-separated-by-conflict-after-16-years-/100413214, August 2021, (21/10/2022).

[7] F. Schroff, T. Treibitz, D. Kriegman, S. Belongie, “Pose, illumination and expression invariant pairwise face-similarity measure via Doppelgänger list comparison”, International Conference on Computer Vision, pp. 2494-2501, 10.1109/ICCV.2011.6126535, 2011.

[8] H. Lamba, A. Sarkar, M. Vatsa, R. Singh, and A. Noore, “Face recognition for look-alikes: A preliminary study”, International Joint Conference on Biometrics (IJCB), pp. 1-6, 10.1109/IJCB.2011.6117520, 2011.

[9] N. L. Segal, J. L. Graham, and U. Ettinger, “Unrelated look-alikes: Replicated study of personality similarity and qualitative findings on social relatedness”, Personality and Individual Differences, vol. 55(2), pp. 169-174, 2013

[10] G. Guo, X. Wang, “Kinship Measurement on Salient Facial Features”, IEEE Transactions on Instrumentation and Measurement, vol. 61(8), pp. 2322-2325, 2012.

[11] M. F. Dal Martello and L. T. Maloney, “Lateralization of kin recognition signals in the human face”, The Association for Research in Vision and Ophthalmology - Journal of vision, vol. 10(8), 2010.

[12] M. F. Dal Martello and L. T. Maloney, “Where are kin recognition signals in the human face?”, The Association for Research in Vision and Ophthalmology - Journal of vision, vol. 6(12), 2006.

[13] J. P. Robinson, M. Shao, Y. Wu and Y. Fu, “Family in the Wild (FIW): A Large-scale Kinship Recognition Database”, CoRR, abs/1604.02182, 2016.

[14] J. Lu, X. Zhou, Y. Tan, Y. Shang and J. Zhou, “Neighborhood Repulsed Metric Learning for Kinship Verification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36(2), pp. 331–345, 2014.

[15] L. M. DeBruine, F.G. Smith, B. C. Jones, S. C. Roberts, M. Petrie and T. D. Spector, “Kin recognition signals in adult faces”, Vision research - Elsevier, vol. 49(1), pp. 38–43, 2009.

[16] G. Kaminski, S. Dridi, C. Graff, and E. Gentaz, “Human ability to detect kinship in strangers’ faces: effects of the degree of relatedness”, Proceedings of the Royal Society B: Biological Sciences, vol. 276(1670), pp. 3193-3200, 2009.

[17] A. Alexandra, P. Fanny, M. Allan, M. Ulrich and R. Michel, “Identification of visual paternity cues in humans”, Biology letters, vol. 10(4), 2014.

[18] L. T. Maloney, and M. F. Dal Martello F., “Kin recognition and the perceived facial similarity of children”, Journal of Vision, vol. 6(10), 2006.

[19] H. Yan, J. Lu, W. Deng, and X. Zhou, “Discriminative Multimetric Learning for Kinship Verification”, IEEE Transactions on Information Forensics and Security, vol. 9(7), 2014.

[20] R. Fang, K. D. Tang, N. Snavely, and T. Chen, “Towards computational models of kinship verification”, Proc. IEEE International Conference on Image Processing (ICIP), 2010, pp. 1577-1580.

[21] X. Qin, X. Tan, and S. Chen, “Tri-subjects kinship verification: Understanding the core of a family”, IAPR International Conference on Machine Vision Applications (MVA), pp. 580-583, 2015.

[22] J. Zhang, S. Xia, H, Pan, and A. K. Qin, “A genetics-motivated unsupervised model for tri-subject kinship verification”, IEEE International Conference on Image Processing (ICIP),pp. 2916-2920, 2016.

[23] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis and Y. Fu, “Visual Kinship Recognition of Families in the Wild”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40(11), pp. 2624–2637, 2018.

[24] R. F. Rachmadi, I. K. E. Purnama, S. M. S. Nugroho and Y. K. Suprapto, “Family-aware convolutional neural network for image-based kinship verification”, International Journal of Intelligent Engineering and Systems, vol 13(6), pp. 20–30, 2020.

[25] F. Schroff, D. Mathematical Problems in Engineering Kalenichenko and J. Philbin, “Facenet: A unified embedding for face recognition and clustering”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[26] L. Dulčić, “Face Recognition with FaceNet and MTCNN - Ars Futura”, https://arsfutura.com/magazine/face-recognition-with-facenet-and-mtcnn/, (Accessed on 21/10/2022).

[27] R. Fang, K. Tang, N. Snavely, and T Chen, “Towards computational models of kinship verification”, IEEE International Conference on Image Processing, pp. 1577-1580, 2010.

[28] G. Guo and X. Wang, “Kinship measurement on salient facial features”, IEEE Transactions on Instrumentation and Measurement, vol. 61(8), 2012.

[29] M. Xu and Y. Shang, “Kinship Verification Using Facial Images by Robust Similarity Learning”, Mathematical Problems in Engineering, pp. 1-8, 2016.

[30] “The Closed Eyes in the Wild (CEW) dataset”, http://parnec.nuaa.edu.cn/_upload/tpl/02/db/731/template731/pages/xtan/TSKinFace.html , (Accessed on 21/10/2022).

[31] J. Liang, Q. Hu, C. Dang, and W. Zuo, “Weighted graph embedding-based metric learning for kinship verification”, IEEE Transactions on Image Processing, vol. 28(3) pp. 1149–1162, 2019.

[32] R. Fang, A. C. Gallagher, T. Chen, and A. Loui, “Kinship classification by modeling facial feature heredity”, IEEE International Conference on Image Processing, pp. 2983-2987, 2013.

[33] T. H. Nguyen, H. H. Nguyen and H. Dao, “Recognizing families through images with pretrained encoder”, arXiv, 2020.

[34] Seldon, “Transfer learning for machine learning”, https://www.seldon.io/transfer-learning/, June 2021, (Accessed on 21/10/2022).

[35] S. Ronaghan, “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark”, Towards Data Science, https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3, (Accessed on 21/10/2022).

[36] “Advantages of a Decision Tree for Classification - Python”, https://pythonprogramminglanguage.com/what-are-the-advantages-of-using-a-decision-tree-for-classification/, (Accessed on 21/10/2022).

[37] S. M. Piryonesi and T. E. El-Diraby, “Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems”, Journal of Transportation Engineering, Part B: Pavements, vol. 146(2), 2020.

[38] H. Ishwaran, “The effect of splitting on random forests”, Springer, vol. 99(1), pp. 75-118, 2015.

[39] S. Nembrini, I. R. König, M. Wright, “The revival of the Gini importance?”, Bioinformatics, vol. 34(21), pp. 3711-3718, 2018.

[40] N. Liberman, “Decision Trees and Random Forests”, Towards Data Science, https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991, (Accessed on 21/10/2022).

[41] “4.2. Permutation feature importance — scikit-learn 1.0.1 documentation”, https://scikit-learn.org/stable/modules/permutation_importance.html, (Accessed on 21/10/2022).

[42] “Naive Bayes Classifier”, Machine Learning Simplilearn, https://www.simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-classifier, (Accessed on 21/10/2022).

[43] “Naive Bayes Classifier: Pros & Cons, Applications & Types Explained”, upGrad blog, https://www.upgrad.com/blog/naive-bayes-classifier/, (Accessed on 21/10/2022).

[44] A. Bakharia, “Visualising Top Features in Linear SVM with Scikit Learn and Matplotlib”, Medium, https://aneesha.medium.com/visualising-top-features-in-linear-svm-with-scikit-learn-and-matplotlib-3454ab18a14d, (Accessed on 21/10/2022)

[45] “SVM: Advantages Disadvantages and Applications”, Statinfer, https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/, (Accessed on 21/10/2022).

[46] “Advantages of Support Vector Machines (SVM)”, https://iq.opengenus.org/advantages-of-svm/, (Accessed on 21/10/2022).

[47] J. Cervantes, X. Li, W. Yu, and K. Li, “Support vector machine classification for large data sets via minimum enclosing ball clustering”, Neurocomputing, vol. 71(4), pp. 611-619, 2008.

[48] scikit-learn 1.0 documentation, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, (Accessed on 21/10/2022)

[49] “Advantages and Disadvantages of Logistic Regression”, https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/, (Accessed on 21/10/2022).

[50] “Advantages and Disadvantages of Logistic Regression - GeeksforGeeks”, https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/, (Accessed on 21/10/2022).

An Assessment of Human Depth Understanding in Handheld Light-Field Displays

Raymond Swannack, Oky Dicky Ardiansyah Prima
Graduate School of Software and Information Science
Iwate Prefectural University
152-52 Takizawa, Iwate, Japan
e-mail: g231s501@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

Abstract—**Light-field displays (LFDs) allow users to view stereoscopic images without the need for a headset, providing a novel 3-Dimensional (3D) experience. This study aims to expand upon the preliminary experiment, in which we evaluated the benefits of stereoscopic cues in human visual understanding through users performing 3D interactions on a multi-view, LFD display. Our task scenario involves user tests for 3D alignment accuracy and a questionnaire about the experience during the test. For each task, using the LFD "Lume Pad" developed by Leia Inc., 3D contents are presented with stereoscopic cues and without. Results from subjects showed that task alignment could be achieved with greater accuracy when stereo cues were available than when there were not. The questionnaire showed that depth perception appeared to be easier to comprehend with stereoscopic cues.**

Keywords-component Light-Field Display; 3D human perception; motion-parallax; stereoscopic vision; head tracking.

I. Introduction

Two-Dimensional (2D) screens have almost limitless possibilities. These displays can show locations that the user has never seen, visualize data of almost any form, and allows professionals to interact with information in ways that are difficult within a physical medium. Even with all these possibilities, the 2D screen is not perfect. This style screen cannot show true depth, as it is a flat object and does not have any depth to it. This paper is an extension of our previous work, evaluating the benefits of depth comprehension of an LFD over a standard flat screen [1].

The human eye does not interact with a screen in the same way as it does with the real 3D world [2], especially in terms of seeing depth. To aid in recreating depth, there are many tools that can be used to simulate depth cues. There are nine widely agreed-upon sources of information that the human brain uses for the purpose of perceiving depth. They are as follows, binocular disparity, convergence, occlusion, relative size, height in the visual field, relative density, aerial perspective, accommodation, and motion parallax [3]. To a greater or lesser extent, these are the primary tools used in conveying the illusion of depth within a 2D screen, and it is the manipulation of these sources that forms the basis of 3D displays. By putting more focus on one source than others, many different types of 3D displays can be created, each being tailored to suit different tasks.

For most of the populace, VR represents a VR headset, or Head Mounted Display (HMD) as these have had the most exposure in popular culture. The popularity of products such as the HTC VIVE and the Oculus sold by Meta are best known for their entertainment uses but are becoming more well known for their benefits to training, such as cancer patient care [4], as well as in scientific research. The basis of VR HMDs is to use two small, high-resolution displays, placed close to the wearer's eyes. Each display is positioned so that only one eye can see each screen, thus using binocular disparity to create a stereoscopic experience. Each screen displays a slightly different view of the same scene, allowing the user's brain to put them together to create a 3D image. This is to simulate how the human eyes naturally work in the real world where human eyes are slightly split apart, giving us two slightly different views.

Hand-held displays are not usually considered to be a type of 3D display, but they are also capable of displaying believable 3D images. These displays are largely composed of devices not specifically designed for this purpose such as smartphones and tablets, though there are some that are specifically designed to be used for 3D content such as the RED Hydrogen One smartphone and the Nintendo 3DS. Many hand-held displays use apps and programs, such as Pokémon Go or IKEA Place, that use the device's inbuilt camera to give the appearance of projecting objects into the real world via the device's screen. This style of software displays what the camera sees and adds digital objects to the scene to show the user a believable 3D scene. In so doing, this type of 3D display relies largely on height in the visual field, comparing the height of the digital objects to the size of known physical objects in the scene, for the user to believe that what they are looking at is real.

Another device that is designed to create stereoscopic images in a different way is known as an LFD. LFDs use curved lenses, such as lenticular lenses, to redirect the light coming out of the screen of the display [5]. By doing this, the LFD can split the display so that it gives a different view to each eye. In this way, an LFD works similarly to an HMD, though the LFD performs this job without the need for a headset. This also allows for more than one user to be able to see the 3D effect from the same display.

The purpose of this study is to improve on our preliminary findings and to show that adding stereoscopy to a tablet display increases a user's understanding of what it is that they are seeing. To this end, the new experiment has more easily understood user controls, so the experiment emphasizes visual understanding and not mastery of the controls. The experiment also has less visual stimulation, to help the subjects focus on the important details of the experiment.

The remainder of the paper is structured as follows. In Section II, we present our experiment's methodology for evaluating the subject's accuracy with and without stereoscopy. Section III covers the details of the hardware and software used in the experiment. Our preliminary experiment is discussed in Section IV, both our findings and what we felt needed to be improved for a follow-up experiment. Section V details the experiment as well as the results and questionnaire. In Section VI we discuss our findings and the implications. Finally, we finish our work in Section VII, with our conclusion and discuss future work.

II. Methods

The focus of this study was to measure a subject's understanding of a 3D scene displayed on a 2D screen, given different visual cues. To measure accuracy, subjects were asked to aim an arrow at a target, using the visual cues available to them to attempt to hit as close to the center of the target as they could. This test was repeated four times with some changes to assess the subject's understanding of the scene.

A. Visual Cues

The primary visual cues that are observed in this test are occlusion, motion parallax, and stereoscopy. Occlusion happens when an object blocks the line of sight to another object. This technique has been used for centuries as a method of showing depth in many different forms of illustrations.

Humans see objects closer to them as moving faster than objects that are further away. This is more prevalent in a vehicle moving at speed and is known as motion parallax. Both visual cues are prevalent throughout the entire experiment.

B. Accuracy Test

To assess the subject's understanding of the simulated depth within the screen, a 3D scene was created to interact with. This scene includes an arrow, target, obstacles, and minute details to enhance the visual cues within the scene.

The target, as seen in Figure 1, is at a set *y* position, 7 meters, and is offset by a distance in the *x* and *z* directions. This distance changes for each test, as seen in Figure 2, altering the ideal arrow angle and camera position for each test.

To obscure the subject's view and encourage them to interact with the scene, obstacles are added to the scene. A rock is placed directly behind the arrow, stopping subjects from simply positioning the camera behind the arrow and lining up their aim this way. By doing this, the subject would not require the visual cues we are evaluating. There are also four trees that are placed to further complicate the scene and encourage more interaction. These trees change their location after each test much like the target.

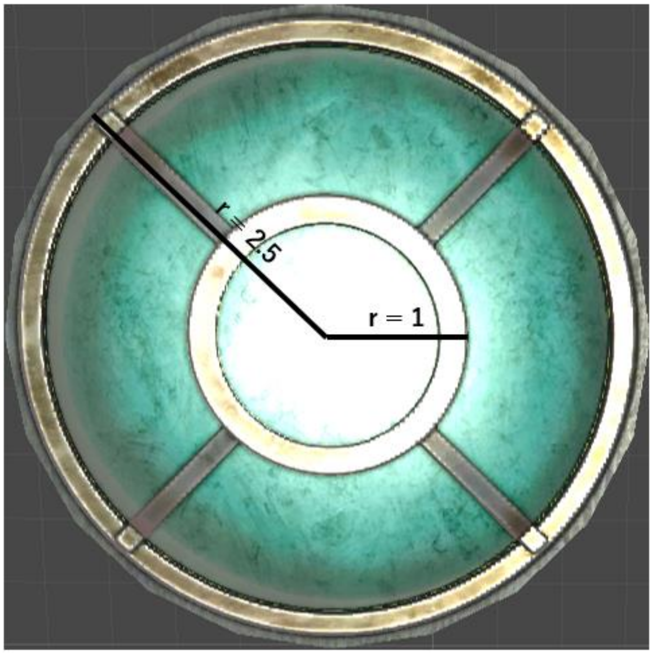


Figure 1. Target.

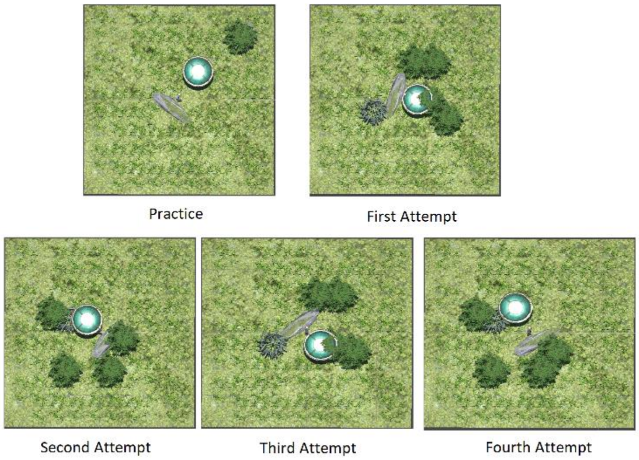


Figure 2. Top-down view of Accuracy Test.

C. Accuracy Measurement

To measure the subject's understanding of the scene, the distance from the arrow to the center of the target is measured. Euclidean distance is used for this measurement. A lower distance is desirable as it shows the subject was able to aim the arrow close to the center of the target.

$$Distance = \sqrt{(x_1 - x_2)^2 + (z_1 - z_2)^2} \quad (1)$$

In this equation, x_1 and z_1 correspond to the position of the target while x_2 and z_2 to the final location of the arrow. The unit of measure for the experiment is Unity units, which are equivalent to meters (m). A score of less than 1 was considered a good score, as it hit within the center ring of the target, while a score between 1 and 2.5 was considered an acceptable score, as it hit the outer ring. A score of above 2.5 missed the target and was considered a poor score.

D. *Head Tracking*

Another key to this research was the decision to perform head tracking. With an LFD, head movements change the view that the user is seeing, therefore moving around gives different viewing angles into the same scene. We believe that subjects will benefit from seeing these different views. Head movement is not required for the test though. A user can perform the entire test without moving their head enough to shift which view they are seeing.

Two VIVE trackers were used to monitor the position of the subject's head in relation to the position of the tablet. One tracker was placed on a stand so that it is positioned just below the tablet. The tracker sits 8 cm above the table, with the bottom of the tablet being 13 cm above the table and the top of the tablet being 30 cm above the table. The second tracker was attached to the user's head via a head strap. This allowed for the user's head position and rotation to be monitored both on their own, as well as in relation to the tablet, as seen in Figure 3.

E. *Experimental Procedure*

The procedure for the experiment was as follows. Firstly, the subject was sat at the table, put on the head tracker, and told what was expected of them as well as given a description and short demonstration of how the controller worked. Then they were given a practice scene and told to spend as much time as they needed to feel comfortable with how to control the arrow and move the camera but were instructed to not press the X button. Once they felt comfortable, they were instructed to press the X button and the test began. The subject performed the first four tests, then the experiment was reset, and the Light-field effect was either turned on or off, depending on which group the subject was in. After this, they were asked to do another four tests. Finally, they filled out a short questionnaire.

Subjects sat 45-55 centimeters from the Lume Pad as seen in Figures 3 and 4. This is the distance that Leia Inc states is the best viewing distance for observing the stereoscopic visual cues that the Lume Pad produces. The tablet itself was positioned on a stand that was adjusted for each user to give them the best viewing angle. The subject's horizontal position was not considered as the subjects were free to move to observe the different views displayed by the LFD.

Subjects were separated into two groups. Group one performed the practice as well as the first four tests without the Light-field effect turned on, then the Light-field effect was turned on and they performed another four tests. Group two performed the practice and first four tests with the Light-field effect turned on, then it was turned off for the second four tests.

III. Hardware and Software used in Experiment

A. *Lume Pad*

The following research was performed using a Lume Pad, an LFD tablet produced by Leia Inc. As discussed above, this allows users to see the illusion of depth inside a 2D screen by creating stereopsis. The tablet boasts a 10.1-inch screen with a resolution of 2560x1600 pixels.

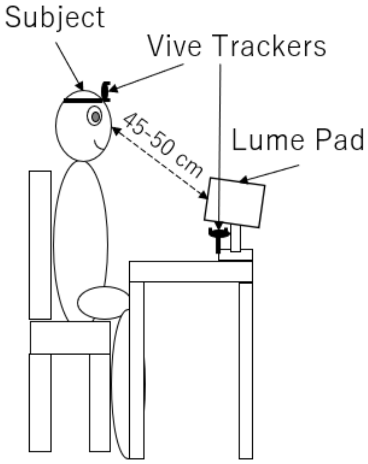


Figure 3. Experiment Setup.

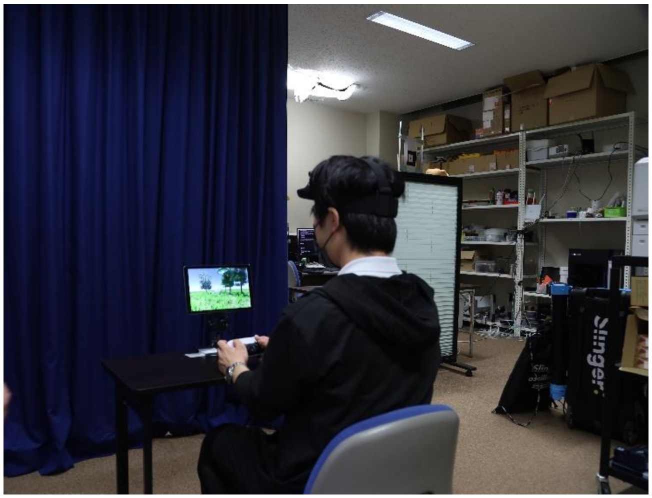


Figure 4. Subject Participating in Experiment.

To create the light field effect, the tablet displays four views at the same time and uses lenticular lenses to allow the user to see two of these images at a time. If the user moves to the sides they will see two different images, thus creating a different view. In this way, the Lume Pad creates three views that can be viewed by changing the user's viewing angle.

To generate the different views, the Lume Pad stores four images internally as a single image file broken into a 2x2 grid. Due to needing to put four images into one screen, each image can only make use of one-quarter of the total resolution, so each view has a resolution of 640x400 pixels [6].

One way to measure how much detail a screen can show is through pixel density. This is calculated as follows,

$$PPI = \frac{Diagonal\ in\ Pixels}{Diagonal\ in\ Inches} \qquad (2)$$

$$Diagonal\ in\ Pixels = \sqrt{Width^2 + Height^2} \qquad (3)$$

The width and height pertain to the dimensions of the tablet. In this case, the Lume Pad has a width of 2560 pixels and a height of 1600 pixels. Given this, the pixel density of each image within an LFD image on the Lume Pad, is 75 pixels per inch (ppi) which is small compared to the potential of the tablet without the Light-field turned on, which has a pixel density of 290 ppi. This is not a favorable comparison as the LFD uses only 25.9% of the pixels per image for each of its images, though this is to be expected and is a misnomer. For a more realistic comparison, the LFD can be compared to a standard computer monitor. A popular computer screen size is 24 inches with a resolution of 1920x1080 pixels, giving this screen a pixel density of 92 ppi. This is a much more favorable comparison for the Lume Pad with the LFD turned on, as it has 81.5% of the pixel density of a standard computer screen.

The Lume Pad uses a few techniques to make the display appear to have a clear picture even with its slightly low resolution. These include having a smaller screen size compared to a desktop or laptop computer as well as the orientation of the lenticular lenses. These lenses are not aligned vertically but are instead slanted slightly. This allows for smoother transitions between views as well as allowing for vertical changes of view and not just horizontal changes of view. It has also been shown as an effective way to blend views together, making the user believe they are able to see more views than are being displayed [7].

B. Sony Dual Sense Controller

A Sony Dual Sense controller was chosen as the input tool for this experiment for a few reasons. Firstly, it is Bluetooth compatible, so it can easily connect to the Lume Pad. Secondly, it is a familiar controller in both shape and layout to many people as it follows a similar layout to popular home gaming systems.

The directional pad (d-pad) on the left side of the controller, is used to control the arrow, allowing the user to aim it to the facing that the subject believes will hit the center of the target to the best of their ability. On the right side of the controller, the southern button, the X, on the controller, launches the arrow. The western button, the Square button, resets the test and the northern button, the Triangle button, turns the LFD effect on and off.

The left stick controls the scene rotation, but it only rotates the scene camera around the center point on the horizontal axis. There is no way to move the camera on the vertical axis. This decision was made because the Lume Pad works best with multiple horizontal views and not multiple vertical views. The right shoulder button zooms the camera in while the left zooms the camera out. The full controller layout can be seen in Figure 5.

C. Unity

Our experiment was designed using the Leia Unity Software Development Kit (SDK) [8]. This SDK allows for the utilization of the Lume Pad's features, such as the special Leia camera, as well as having the ability to turn the LFD on and off within the test. The Leia camera is four cameras, aligned in parallel with each other. Aligning the cameras in parallel is important to avoid the keystone distortion and depth plane curvature, as seen in Figure 6, that can occur with a toed-in camera [9]. This distortion is not as visible at close range but at a longer range the image warps in a way that does not align with how human eyes naturally see the world.

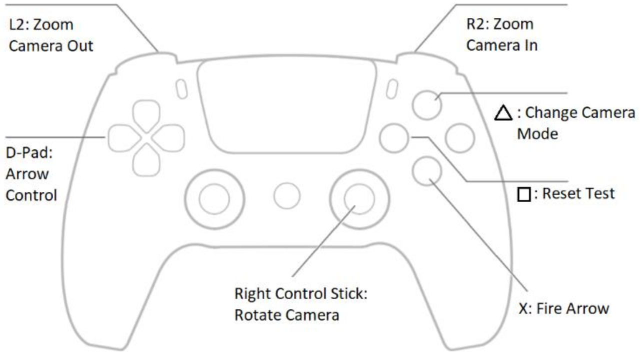
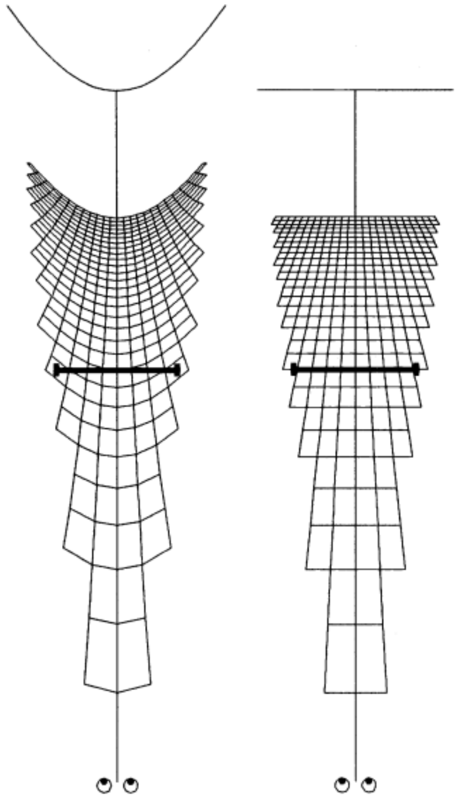


Figure 5. Sony Dual Sense Controller - Layout.



(a) Keystone Distortion (b) No Distortion

Figure 6. Keystone Distortion.

IV. Preliminary Experiment

In the preliminary experiment, 12 subjects (3 females and 9 males) were asked to fire an arrow at a target. Subjects were split into two groups, with group one testing with the Light-field turned off first and then turned on, while group two did the same but in reverse. In that experiment, the subjects performed three tests, with the target being placed further away from the arrow with each test. Then the LFD was switched off or on, depending on the subject's group, and three more attempts were performed. The data collected was the distance from the center of the target as well as the time it took for each attempt.

A. Results

The results can be viewed in Table 1. The scores are the median value of all attempts that each subject performed. The overall mean of the tests with the Light-field turned on, 1.881, was lower than when it was turned off, 2.264. This shows that overall, the subjects were on average more accurate with the Light-field turned on. Furthermore, we can see from this data that group two was more accurate than group one on average. Within the groups, another interesting point can be observed. Group one was more accurate on average with the Light-field effect turned on than with it off, with scores of 2.119 and 2.666, respectively. Group two follows the same pattern, with scores of 1.644 with the Light-field turned on and 1.863 with it turned off.

While it is tempting to state that this proves our hypothesis that stereoscopy increases the user's understanding and thus their accuracy, this data does not allow us to state this. A two-way ANOVA to analyze the effect of test order and the Light-field effect on subject accuracy was performed. Simple main effects analysis showed that test order did not have a statistically significant effect on subject accuracy (p = 0.1118). Similarly, the analysis showed that the Light-field effect did not have a statistically significant effect on subject accuracy either (p = 0.3305).

The two-way ANOVA revealed that there was not a statistically significant interaction between the effects of test order and the Light-field effect ($F(1, 20) = 0.1829$, p = 0.6735).

These results show that the preliminary experiment strengthens the hypothesis that the stereo effect of the LFD was beneficial to the subjects, but the results were not conclusive. We observed that the first group did get more accurate when stereoscopy was added and that the second group was more accurate while using stereoscopy than without. The results of the ANOVA were a large problem though, as it is not clear that it is mainly the Light-field effect that is affecting subject accuracy.

B. Improvements for the Main Experiment

The most critical issue that needed addressing in further experiments was the precision of the controls. Some of the subjects stated that the controls were too imprecise to feel confident in their accuracy with these tests. Trying to make minute changes to the aim of the arrow proved difficult to achieve using the dual sense controller's control stick. For some subjects, this was partially due to unfamiliarity with the controller. A control stick is intuitive to those that often play video games, but confusing to those who do not.

Another reason is due to how hard it is to control an object when it has free movement in three dimensions. While some users were able to control the arrow without much effort, feeling that they understood where the arrow was pointing at all times, others constantly overcorrected their aim and had to try to bring it back to where they wanted the arrow pointing. This tested some subjects' patience, and a few seemed to decide they were close enough instead of trying to be as precise as possible. By making the controls more precise and easier to operate, we believe that users would be more confident in their accuracy with the test, and as such it would be clearer how much of a significant role the Light-field played in subject accuracy.

Based on feedback and further research, changes were proposed to the scene itself. Moving the target closer to the arrow is one such change. In the preliminary experiment, the later tests placed the target at a distance that made it difficult to understand where it was on the small screen.

Some scenery was removed as well, being one of the trees and a portion of the grass. The tree was deemed to be unnecessary as there were already enough obstacles blocking sight lines. The grass was there to give the user more visual cues to perceive motion parallax, but the amount of detail that the display needed to render was affecting its performance and we also worried that there were perhaps too many cues for the user [10].

In our previous paper, we discussed performing eye tracking as a possible addition to future research, but it was decided that head tracking would give more significant data than eye tracking. The reasoning behind this was that eye tracking would show eye vergence [11], giving a better understanding of how subjects' brains perceive the scene displayed on the LFD, while head tracking indicates the user's interaction with the device. Head tracking was chosen over eye tracking because an LFD is a flat screen, so eye vergence will not be the same as if the scene was in real 3D and would likely behave as they were interacting with a standard 2D screen. Thus, head tracking was deemed to be more noteworthy.

TABLE 1. PRELIMINARY EXPERIMENT RESULTS

		Subject	Display		Mean	St. Dev
			Without LFD	With LFD		
Group	I	1	2.234	2.175	2.392	1.2167
		3	2.054	1.673		
		5	2.408	1.993		
		7	2.490	1.907		
		9	5.422	4.058		
		11	1.387	0.905		
	II	2	1.363	1.859	1.753	0.4729
		4	1.580	0.861		
		6	1.927	1.946		
		8	2.842	2.083		
		10	1.731	1.433		
		12	1.733	1.681		
		Mean	2.264	1.881		
		St.Dev	1.0935	0.8069		

V. Experiment and Results

For the experiment, there were 8 subjects (one female and seven males). They ranged in age from 22 to 27 with an average age of 25. All had either normal or corrected-to-normal vision. The experiment was carried out over two days, with subjects asked not to discuss the test with other subjects. Each subject was assigned a number, with odd-numbered subjects placed in group one while even-numbered subjects were placed in group two.

Table 2. Main Experiment Results

		Subject	Display		Mean	St. Dev
			Without LFD	With LFD		
Group	I	1	0.318	0.519	1.229	1.0047
		3	1.321	0.722		
		5	3.454	1.640		
		7	0.641	1.218		
	II	2	1.467	0.742	1.511	0.8159
		4	0.462	0.912		
		6	1.998	1.343		
		8	2.748	2.412		
		Mean	1.551	1.189		
		St.Dev	1.1244	0.6174		

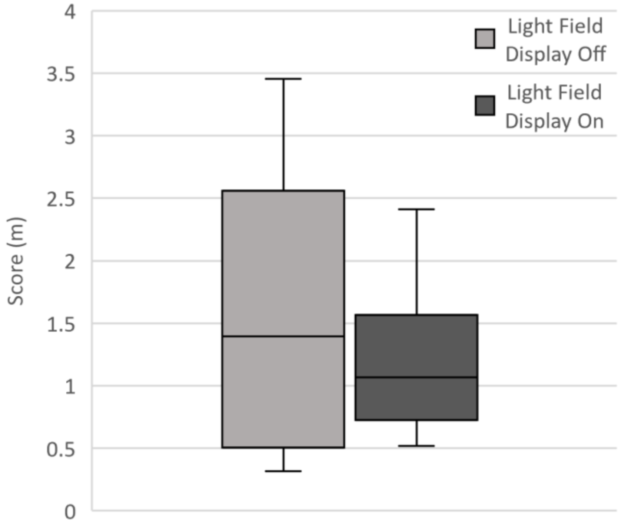


Figure 7. Interquartile Range

A. Scores

Table 2 displays the accuracy results of the experiment. From this table, it can be observed that on average, subjects were more accurate when they had access to stereoscopy, with the LFD effect turned on. On average with the LFD turned on, the average score was 1.189, which would miss the inner circle of the target by 0.189 meters. The low variance, the square of the standard deviation, of 0.381 means that this is a rather reliable estimation. With the LFD turned off, the subjects on average had a score of 1.189, which would hit the outer circle of the target, though with a high variance of 1.264 this number is not as reliable as the results with the LFD on.

Due to the small data set, another way to evaluate this data is by looking at the interquartile range. This is a look at the range of the inner two-quarters of the data. It is useful in small data sets by removing outliers and observing the distance between the data around the mean. When the LFD is turned off the interquartile range is 1.826, while with the LFD turned on the interquartile range is 0.760. Figure 7 visualizes this information.

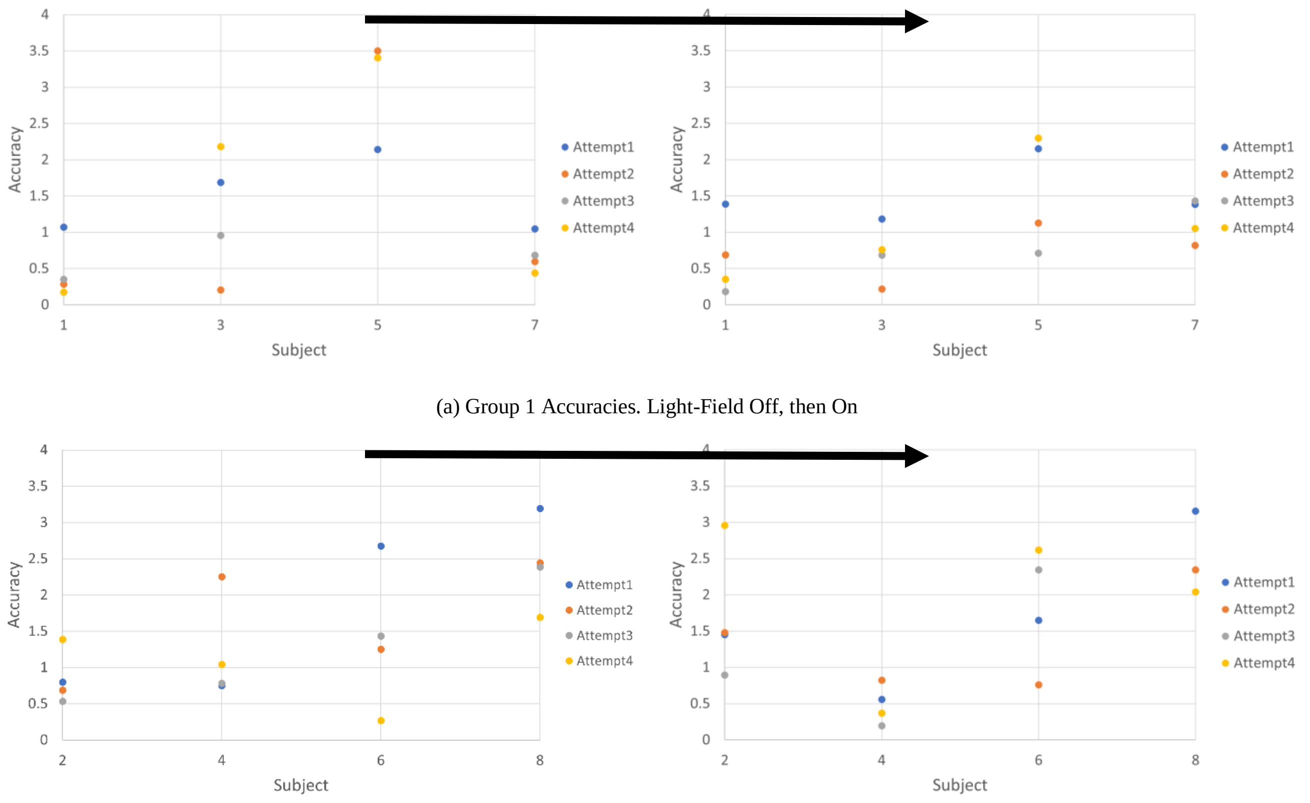
As with the preliminary experiment, a two-way ANOVA was performed to analyze the effect of test order and the Light-field effect on subject accuracy. Simple main effects analysis showed that test order did not have a statistically significant effect on subject accuracy (p = 0.6944). Similarly, the analysis showed that the Light-field effect did not have a statistically significant effect on subject accuracy either (p = 0.4676).

The two-way ANOVA revealed that there was not a statistically significant interaction between the effects of test order and the Light-field effect (F(1, 12) = 0.0091, p = 0.9256).

Figure 8 allows for a more visually direct comparison of the individual results of the two groups and the two tests. Within Group 1, shown in Figure 8 (a), subjects 3 and 5 both were more accurate when the Light-field was turned on for the second half of the experiment. Subjects 1 and 7 were less accurate, but both were still very accurate and did not miss the target.

Group 2 is similar, as shown in Figure 8 (b), where subjects 2 and 8 are less accurate when the Light-field is turned off, subject 8's third test is so far off the target it does not appear on the graph. Subject 4 is more accurate when the Light-field it turned off, and subject 6 appears to be the roughly the same over both tests.

Furthermore, it can be seen that the target was missed twice with the Light-field effect turned on but was missed seven times with it turned off. The center ring was hit 14 times in both tests.



(a) Group 1 Accuracies. Light-Field Off, then On

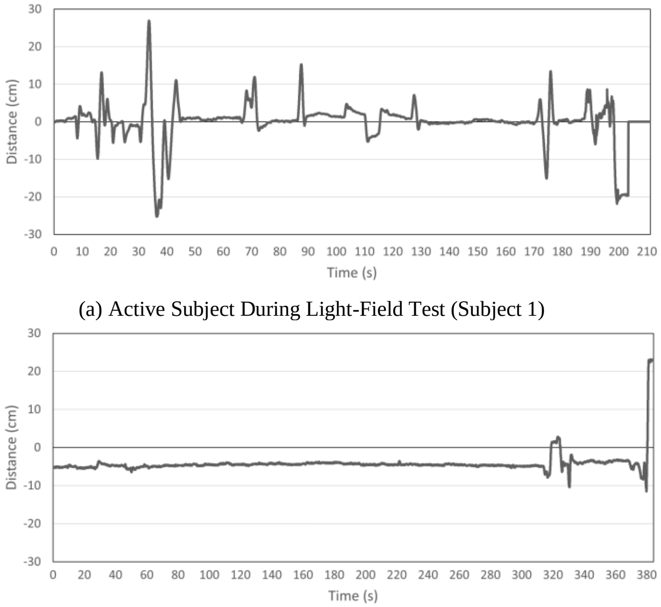
(b) Group 2 Accuracies. Light-Field On, then Off

Figure 8. Accuracy Plot

B. *Head Tracking*

The head tracking data can be observed in Figures 9 (a) and (b). As only the horizontal distance changes the view that the subject sees within the LFD, the X value is the only variable that we have recorded in these graphs. Within the graphs, the distance, measured in centimeters, to the left or right of the centerline of the tablet is displayed. The distance traveled to the right is recorded as a positive value while the distance traveled to the left is recorded as a negative value.

A comparison of subjects 1 and 7 shows the difference in how each interacted with the tablet. Subject 1 scored the best median accuracy over the Light-field tests, with a score of 0.519. They actively moved their head while the Light-field effect was turned on, as seen in Figure 9 (a), moving close to thirty centimeters to either side of the centerline of the tablet at times. In comparison, subject 7, whose data can be seen in Figure 9 (b), tried a small head movement at the beginning and then moved very little until the end of the experiment. They had a median score of 1.218. While head movement may not have been the leading factor for this disparity in the two subject’s scores, it could possibly have affected them.



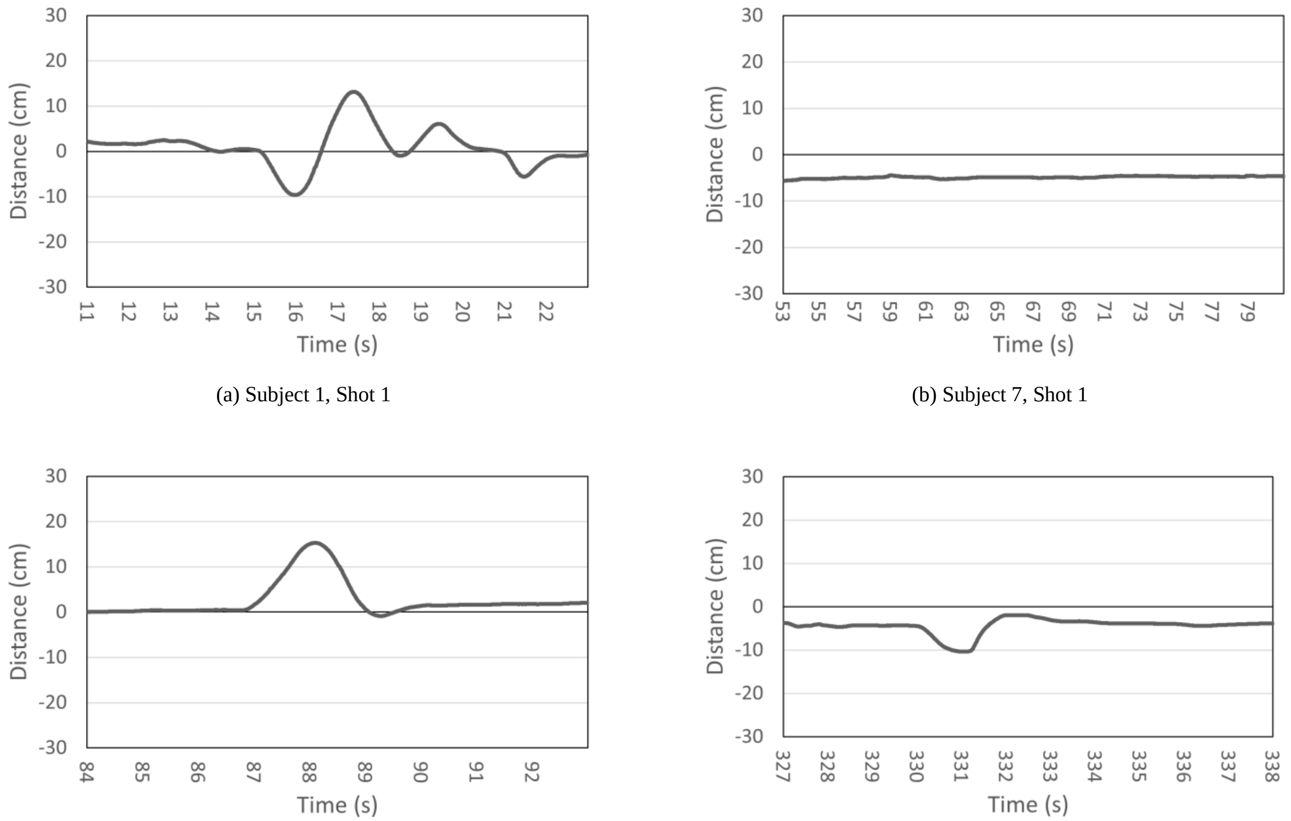
(a) Active Subject During Light-Field Test (Subject 1)

(b) Inactive Subject During Light-field Test (Subject 7)

Figure 9. Comparison of Active and Inactive Subjects

C. *Time Comparison*

Table 3 shows the median time spent by each subject for each test. Some subjects were quicker while others took more time. A further comparison of subjects 1 and 7, specifically looking at the times in which they were performing the Light-field tests illustrates how much time some subjects felt they needed for each portion of the experiment. Overall, Subject 2 spent around 200 seconds performing this portion of



(a) Subject 1, Shot 1

(b) Subject 7, Shot 1

(c) Subject 1, Shot 4

(d) Subject 7, Shot 4

Figure 10. Head Movement Comparisons.

TABLE 3. TIME COMPARISON

		Subject	Display		Mean	St.Dev
			Without LFD	With LFD		
Group	I	1	22.91	24.90	41.29	23.870
		3	73.57	56.07		
		5	19.51	12.26		
		7	54.89	66.18		
	II	2	12.26	24.90	30.52	19.858
		4	23.60	29.10		
		6	66.18	56.07		
		8	16.94	15.13		
		Mean	36.23	35.58		
		St.Dev	24.503	20.729		

the experiment while Subject 7 spent around 350 seconds, over two minutes longer. For their first shot, Figure 10 (a), with the Light-field turned on, subject 1 spends a few seconds observing and interacting with the test before moving their head to look at the different views. At this point, they move 10 centimeters to the left, then about 13 to the right. Afterward, they make much less drastic movements as they finish aiming the arrow. The process takes about 12 seconds. By contrast, in the exact same scenario, Figure 10 (b) shows that subject 7 spends the entirety of their first shot focusing on the scene and not utilizing the other views available if they moved their head. Subject 7 took more than double the time that subject 1 took, at 27 seconds. A similar comparison can be made over their last shot. Subject 1 only makes one movement of around 15cm at about halfway through this shot. They take roughly 9 seconds to perform their final test. Subject 7 by contrast makes a similar movement four seconds into their final attempt which lasts 11 seconds. In both cases, subject 1 spends less time, and moves more. One interpretation of this is that subject 7 spent that time making up for not seeing the other views that Subject 1 utilized.

D. *Questionaire*

Subjects were also given a questionnaire to ascertain how well they believed they had understood the experiment. The questions were as follows:

1) *How confident were you that you would hit the center of the target with the LFD turned off?*

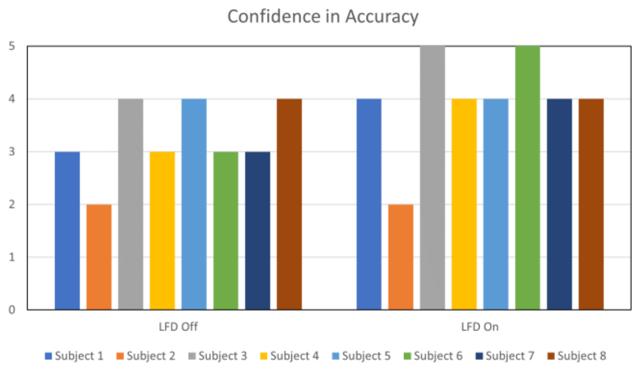


Figure 11. Confidence in Accuracy.

2) *How confident were you that you would hit the center of the target with the LFD turned on?*
3) *Were you more confident with the LFD turned on, off, or neither?*
4) *When the LFD was turned on, did you feel that moving your head to see the other viewing angles helped to improve your accuracy?*
5) *Did you feel dizzy or sick at any time during the test?*

Figure 11 shows the results of the first two questions of the questionnaire. Subjects were asked to rank their confidence on a scale of 1 to 5, where 1 was not confident at all, 3 was somewhat confident, and 5 was extremely confident. Most subjects stated that they felt more confident with the Light-field turned on. Although subjects 2, 5, and 8 ranked their confidence the same over both tests, they stated that they felt they were marginally more confident with the Light-field turned on than with the Light-field effect turned off.

All subjects agreed that they did not feel that moving their head to see other viewing angles helped them. Finally, two subjects felt dizzy when the Light-field effect was turned off, though not so bad that they felt they needed to stop the experiment.

VI. Discussion

Although subjects were encouraged to move their head to see the other views that the LFD generates, every one of them stated that they did not believe that doing this helped with their accuracy. While objects in the foreground and background would move as the view was shifted, the objects near the arrow and the target rarely did. The foreground movement was sometimes slightly helpful, but not enough to make a difference in user confidence.

After further testing, it was concluded that the convergence distance was too short for the different views to have noticeable changes. The convergence distance is the distance from the camera where all camera points converge. One way to think about this is that whatever objects are set at the convergence distance are the focal point. In this test, that was set to be the arrow so that it was always in focus. In the preliminary experiment, the convergence distance was set at 10 meters from the camera, and this led to 2 subjects saying that they felt they could not pull the camera very far from the arrow because then they were unsure of the direction that the arrow was facing. Because of this, both of them took what they believed to be too much time aiming the arrow for each test.

A compromise would be to set the convergence point some distance behind the arrow. This would create more differences in the views. This causes the arrow to be less in focus, making it more difficult to be confident in knowing where the arrow is pointing exactly. With more testing, a good compromise can be achieved.

Another option would be to increase the distance between each camera within the Leia camera. This causes a larger difference between each view from the Lume Pad, and also increases the feeling of stereopsis in the user, as the view given to each eye is spaced further apart. This is also dangerous though as if this is spread too far, it appears unnatural and can cause motion sickness.

VII. Conclusion and Future Work

In this study, we examined the accuracy of a user's understanding, given some constraints, with a 3D scene on an LFD in an attempt to show that the human brain understands a scene displayed on an LFD better than a standard 2D screen. While the subjects were more accurate with the Light-field turned on, it is not conclusive that the human brain understands simulated distance in the light field display more so than that simulated with a 2D screen. With the two-way ANOVA once again being inconclusive we are not able to say definitively that the reason that the subjects were, on average, more accurate with the Light-field on than with it turned off.

To further this experiment, enhancing the change in views when the subject moves their head is the priority. Creating a stereoscopic viewing experience without the use of a headset is the most well-known feature of an LFD, but the fact that it can be viewed from multiple angles to see different views is a feature that should be further explored. Both changing the convergence distance as well as experimenting with increasing the distance between cameras should be fully explored in further work.

Another feature that goes along with the LFD being able to show multiple views at one time is that it can be used by multiple people at the same time. This concept has been explored on large-scale LFDs, such as the 120-degree viewing angle LFD designed by Liu et al. [12]. This LFD is far more complicated than the Lume Pad, requiring vast amounts of space and hardware to set up. The Lume Pad has a much smaller optimal viewing angle but more than one user can still fit within this space. Finding a way to test how well two subjects can interact with and understand what is being shown to them with one Lume Pad would give us useful information for future work.

An idea that would utilize this concept would be a cube-style display. The pCubee, by Stavness et al. [13] is a multiscreen display comprised of 5 screens arranged in a cube formation, each connected to a control board. By using a head tracker and an accelerometer, the control boards change the view on each screen to create a believable 3d image without using stereopsis. Since the cube reacts to the user's interactions, it feels like the user is looking into the cube and seeing real depth. By using Lume Pads, or a similar device, a similar experience could be created, without the need for the head tracker. Similar to the pCubee, an accelerometer would detect movements and inform the screens to change their views accordingly, while the lenticular lenses would display the extra views required for any head movements that the subject performs. It could also be designed to give multiple viewers the same experience at one time. Instead of directing all of the screens to one viewpoint, multiple viewpoints can be defined and the LFDs can create them at the same time.

REFERENCES

[1] R. Swannack and O. D. A. Prima, "Assessment of Differences in Human Depth Understanding Between Stereo and Motion Parallax Cues in Light-Field Displays," The Fifteenth International Conference on Advances in Computer-Human Interactions, ACHI 2022, IARIA, pp. 18-21, 2022.

[2] K. Kato and O. D. A. Prima, "3D Gaze Characteristics in Mixed-Reality Environment," eTELEMED 2021 : The Thirteenth International Conference on EHealth, Telemedicine, and Social Medicine, IARIA, pp.11-15, 2022.

[3] J. E. Cutting and P. M. Vishton, "Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual use of Different Information about Depth," In Perception of Space and Motion, W. Epstein, S. Rogers, Eds. Academic Press, pp. 69-117, 1995.

[4] A. Chirico, et al. "Virtual Reality in Health System: Beyond Entertainment. A Mini-Review on the Efficacy of VR During Cancer Treatment," Journal of Cellular Physiology, 231(2), pp. 275-287, doi:10.1002/jcp.25117, 2015.

[5] P. Benzie, et al. "A survey of 3DTV displays: techniques and technologies," IEEE Transactions on Circuits and Systems for Video Technology, 17(11), pp. 1647-1658, 2007. doi:10.1109/TCSVT.2007.905377.

[6] 3D Lightfield Experience Platform, https://www.leiainc.com/ [Retrieved at May 2022].

[7] C. Van Berkel, and J. A. Clarke, "Characterization and Optimization of 3D-LCD Module Design," Stereoscopic Displays and Virtual Reality Systems IV, 3012, pp. 179-186. SPIE, 1997. doi:10.1117/12.274456.

[8] SDK and Developer Resources, www.leiainc.com/sdk. [Retrieved on 10 June 2022].

[9] A. J. Woods, T. Docherty, and R. Koch, "Image Distortions in Stereoscopic Video Systems," Stereoscopic Displays and Applications IV, 1915, pp. 36-48, SPIE, 1993. doi:10.1117/12.157041.

[10] T. S. Murdison, G. Leclercq, P. Lefèvre, and G. Blohm, "Misperception of Motion in Depth Originates from an Incomplete Transformation of Retinal Signals," Journal of Vision, 19(12), pp. 21-21, 2019. doi:10.1167/19.12.21.

[11] C. Elmadjian, P. Shukla, A. D. Tula, and C. H. Morimoto, "3D gaze estimation in the scene volume with a head-mounted eye tracker," Proc. Workshop on Communication by Gaze Interaction, pp. 1-9, 2018. doi:10.1145/3206343.3206351.

[12] B. Liu, et al. "Time-multiplexed light field display with 120-degree wide viewing angle," Optics Express, 27(24), pp. 35728-35739, 2019. doi:10.1364/OE.27.035728.

[13] I. Stavness, B. Lam, and S. Fels, "pCubee: A Perspective-Corrected Handheld Cubic Display," Proc. SIGCHI Conference on Human Factors in Computing Systems pp. 1381-1390, 2010. doi:10.1145/1753326.1753535.

A Study on the Quality of Facial Expression in Digital Humans

Shiori Kikuchi, Hisayoshi Ito, Oky Dicky Ardiansyah Prima
Graduate School of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate, Japan
e-mail: g231u017@s.iwate-pu.ac.jp, {hito, prima}@iwate-pu.ac.jp

Abstract—**There have been many examples recently of the use of digital humans for interactive communication, such as in customer support and digital health care. The use of digital humans is expected to reduce communication barriers caused by differences in personal appearance, behavior, and facial expressions. Despite the potential for digital humans to represent realistic nonverbal communication, there is a lack of evidence regarding the extent to how effective they are in communication. This study attempts to evaluate the six basic emotions (anger, fear, disgust, happiness, sadness, and surprise) of digital humans with the following two experiments. First, the quality of facial expression is evaluated by an actor's facial expression and the expression of the digital human created from the actor. Second, we evaluate the quality of the facial expressions of 17 subjects captured from various angles and those of digital humans created from the corresponding angles. Two deep learning-based facial expression recognitions: DeepFace and HSEmotion were used for the evaluation. Experimental results showed that HSEmotion demonstrated a more stable recognition rate than DeepFace for the same facial expressions of subjects captured from different directions. However, when the facial expressions of these subjects were transferred to the digital humans, both tools failed to properly recognize their facial expressions. Future work will include a facial expression recognition library that considers both real people and digital humans.**

Keywords-expression; digital human; facial expression; basic emotions; avatar.

I. INTRODUCTION

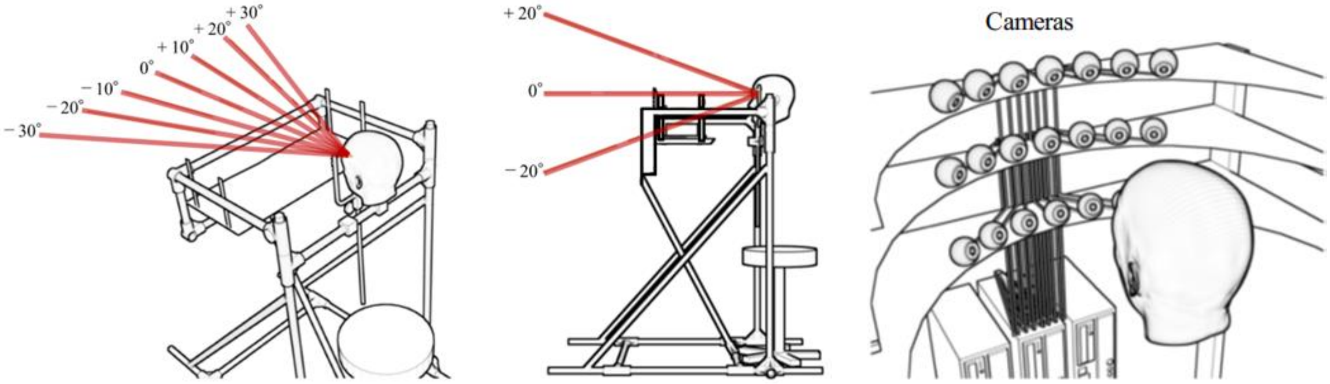
There have been changes in communication methods for forming interpersonal relationships due to the spread of the new corona virus (COVID-19) outbreak, including the use of videoconferencing for meetings. In addition to voice messages, video calls convey visual and non-verbal information, making more effective communication possible. Recent video calls allow participants control over the content to be delivered, such as the ability to remove the video background and manipulate their faces. It is even possible to make video calls using avatars or digital humans, making it easier to communicate with others without worrying about their own appearance. However, facial expressions are not always conveyed adequately by these tools. This paper extends our previous work evaluating facial expressions by digital humans [1].

Significant emotional expression is necessary for communication, which is especially evident in human facial expressions. The quality of digital human facial expressions is therefore considered important for communication in virtual space as well. Interest for communication in these spaces is growing worldwide as well as in Japan, such as "Virtual Shibuya" by KDDI Corporation [2] and "Medical Metaverse Joint Research Chair" by IBM Japan and Juntendo University [3].

Communication in virtual conference applications and games is conducted in real time using virtual characters that have been designed to look like real people in appearance. With the development of artificial intelligence and computer vision, facial expression recognition from facial images is becoming more practical. By transferring the recognized facial components of the user to the virtual character, the user can play various roles through the virtual character. This character is expected to stimulate communication in the medical and business fields.

Digital humans are more realistic than virtual characters, and many are being used in business to enable natural conversations with customers. In addition, digital humans are also equipped with the ability to speak multiple languages, which further expands their applications [4]. Digital humans can be generated from Three-Dimensional (3D) human pose information obtained by capturing a person with a monocular depth camera or a stereo camera [5]. In addition, the development of 3D human pose using a monocular camera, such as the MediaPipe library [6], has become popular, making it possible to create a digital human using only a webcam.

Recently, software tools have been developed to enable the expression of detailed facial expression changes in digital humans using actors' expressions. MetaHuman by Epic Games provides a framework for creating realistic human characters [7]. This framework works by transferring 3D human pose information from the Motion Capture (MoCap) device to the digital human. Similar frameworks include "Character Creator 4" (Reallusion) [8] and "Buddy Builder" (Hologress) [9]. Both frameworks offer a wide variety of resources for 3D clothing and accessories. Buddy Builder features real-time cross-physics, which allows for more natural-looking moving characters. These frameworks are expected to enable sophisticated digital humans to express nonverbal information accurately, which is currently the subject of further research and development. However, the extent to which the quality of digital human facial expressions is comparable to that of humans has not been fully verified.



(a) Horizontal angles (*yaw*). (b) Vertical angles (*pitch*). (c) Location of 21 cameras for capturing.

Figure 1. Multi-angle camera device used in this study.

The purpose of this study is to evaluate the quality of human and digital human facial expressions. The following two attempts were made for the evaluation. First, a digital human that resembles an actor was created, and both facial expressions were compared. Second, basic facial expressions by several subjects were captured using multi-angle camera devices, and these facial expressions were transferred to the digital human. For the evaluation of facial expressions, existing deep learning-based libraries for facial expression recognition are used.

The rest of this paper is organized as follows. Section II discusses the related works of facial expression analyses and digital human. Section III describes the generation of facial expression data and tools used for this purpose in this study. In Section IV, we describe our experiments to evaluate the quality of facial expressions for the digital human. Finally, Section V summarizes the results of this study and discusses future perspectives.

II. Related Works

A. Facial Expressions in Communication

The understanding of one's emotions is important in communication. Particularly, since facial expressions represent human emotions, smooth communication can be achieved by being aware of changes in the facial expressions of conversational partners. Facial expressions consist of movements of small muscles in the face that are used to infer a person's emotional state.

Emotions are difficult to measure since they are often fleeting, hidden, and conflicted. Ekman et al. proposed the Facial Action Coding System (FACS), which classifies Action Units (AUs) of facial parts to identify emotions from facial expressions [10]. Their work pointed out that there are "display rules" based on cultural norms, making the intensity of facial expressions differs from culture to culture. The Japanese, for example, tend to suppress facial expressions [11]. Ohta et al. observed facial expressions in Japanese nursing practice and concluded that the results were generally consistent with Ekman's analysis, but that differences were observed in the distinctness of facial expressions due to the way Japanese people move their facial muscles and their ability to express themselves being weaker than Westerners [12]. Baltrušaitis et al. developed OpenFace [13], a toolkit that can recognize AUs in real time based on facial landmarks taken from the user's face.

B. Facial Expression Recognition based on Deep Learning

As the applications of facial expression recognition have expanded, the development of deep learning-based facial expression recognition has progressed rapidly. With facial expression recognition, an individual's emotional state can be predicted from the appearance of facial deformations. Furthermore, these techniques enable real-time analysis of facial expressions.

Many studies have been conducted in the field of facial expression recognition to realize feasible tasks. In addition, lightweight models have been proposed to enable recognition in web applications and mobile devices. Serengil et al. developed Deepface, a lightweight model of facial expression recognition [13]. Andrey et al. developed a model that utilizes several eEfficientNet-based models to classify emotions of static facial images [14]. This model has been published as the HSEmotion (High-Speed Face Emotion Recognition) [15].

C. Reality of the Digital Human

Digital humans have recently been developed that can resemble humans by appearance and can reflect their body movements [4]. Kang et al. surveyed and simplified the research on digital human reality, introducing two types of reality: visual realism, which is the similarity between the rendering of visual information of a person, and behavioral realism, which is the similarity between human behavior and the reality of a person [16]. Visual realism is high as the digital human looks more like a person, and behavioral realism is high as the digital human performs natural movements. They also stated the importance of the influence of digital human reality on communication. Grewe et al. compared the reality of facial expression animations created by experts with the reality of facial expression animations created statistically

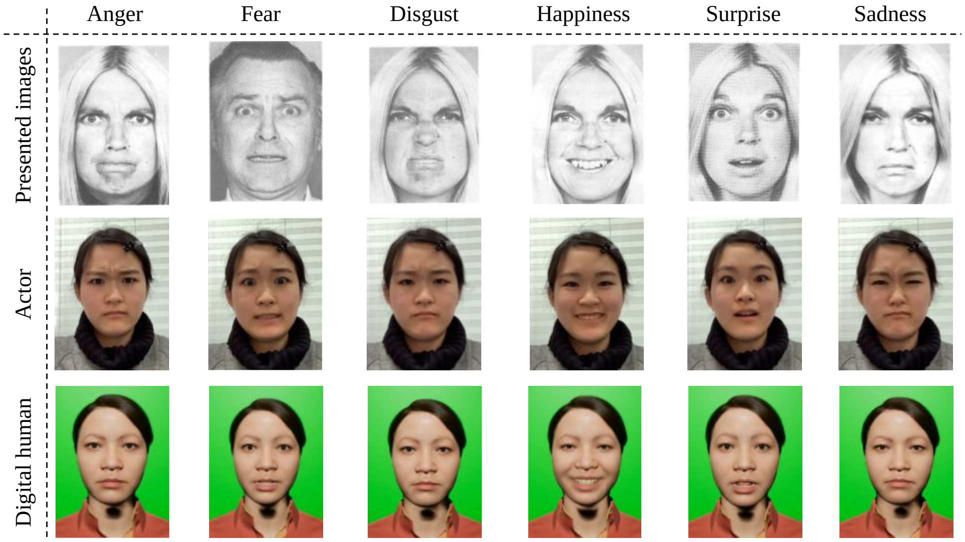


Figure 2. The actor and the generated digital human facial images for each expression.

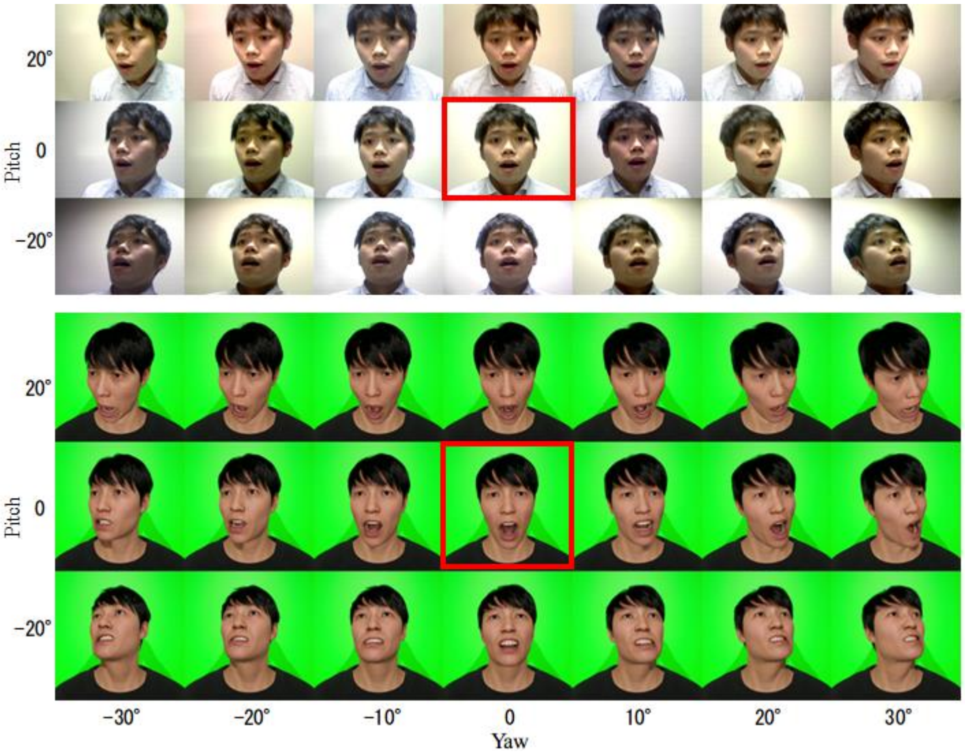


Figure 3. MACD-captured subjects' faces expressing "Surprise," and the digital humans corresponding to each face in each capture direction.

from a database of images. They found that the statistically created animations were perceived as more realistic [17].

III. Generation of Facial Expression Data

This study generated human digital facial expression data using two different capturing methods: frontal and multi-angle photography and evaluated them respectively. Research collaborators photographed in the dataset gave us permission to use the dataset in this study. The methods and tools used to create the dataset and the details of the dataset created are described as follows.

A. Tools Used in This Study

MetaHuman Creator (MHC) and Unreal Engine (UE) by Epic Games were used to create digital humans and reflect the actor's facial expressions on the digital human [18]. MHC is a free cloud-based tool that facilitates the creation of photorealistic digital humans and can be used in conjunction

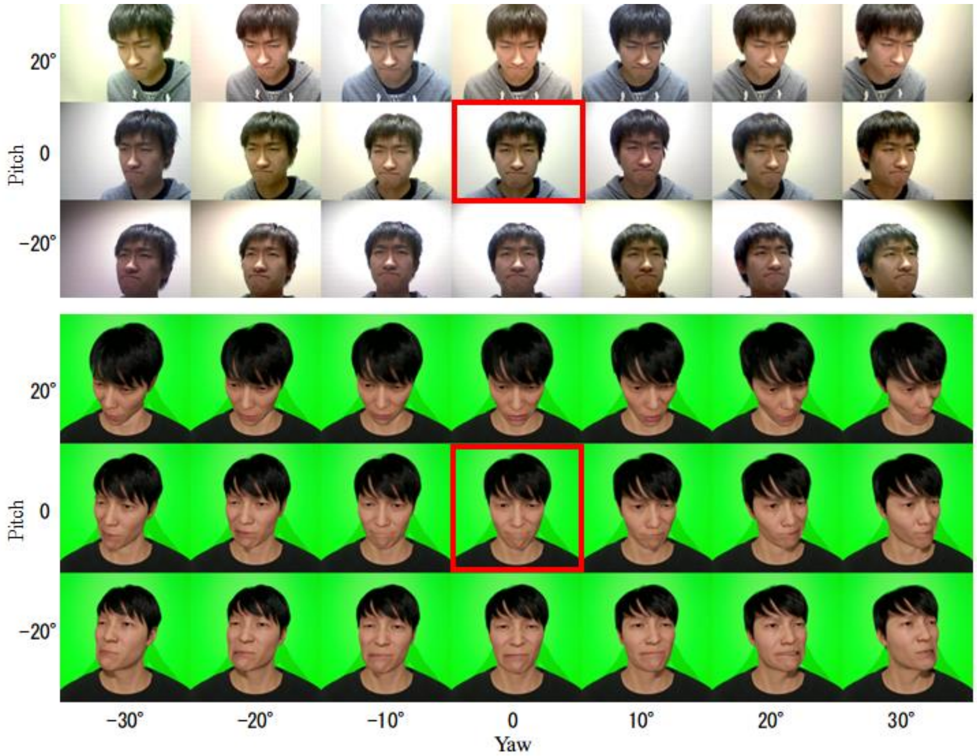


Figure 4. MACD-captured subjects' faces expressing "Disgust," and the digital humans corresponding to each face in each capture direction.

with the 3D object rendering engine of UE. Users can create facial shapes, hairstyles, clothes, and other accessories easily, all of which can be manipulated intuitively through the Graphical User Interface (GUI). This allows average users to create realistic digital humans like computer graphics designers. In addition, MetaHuman data such as meshes, skeletons, facial rigs, animation controls, and materials can be downloaded and exported to other CG software.

MoCap is needed to transfer small facial muscle movements to the MHC. Live Link Face (LLF) and MeFaMo [19] were employed as the MoCap in this study. The former is an iOS application developed by Epic Games that uses the mobile device's depth sensor to extract facial features. The latter uses a monocular camera to estimate and extract facial features based on the MediaPipe library. MeFaMo is used to extract facial muscles in face images obtained from multiple cameras, described in the next subsection.

B. *Multi-Angle Camera Device (MACD)*

In this study, a MACD was constructed to capture facial images simultaneously from multiple angles. Figure 1 shows the device consisting of 21 webcams (640 x 480 pixels, 30 fps). Each camera is arranged in three rows and seven columns. These cameras are networked which allow to simultaneously capture the subject's face at ±30° horizontally (*yaw*) and ±20° vertically (*pitch*) in 10° and 20° increments, respectively. 21 face facial images can be obtained in a single capture.

C. *Facial Expression Data*

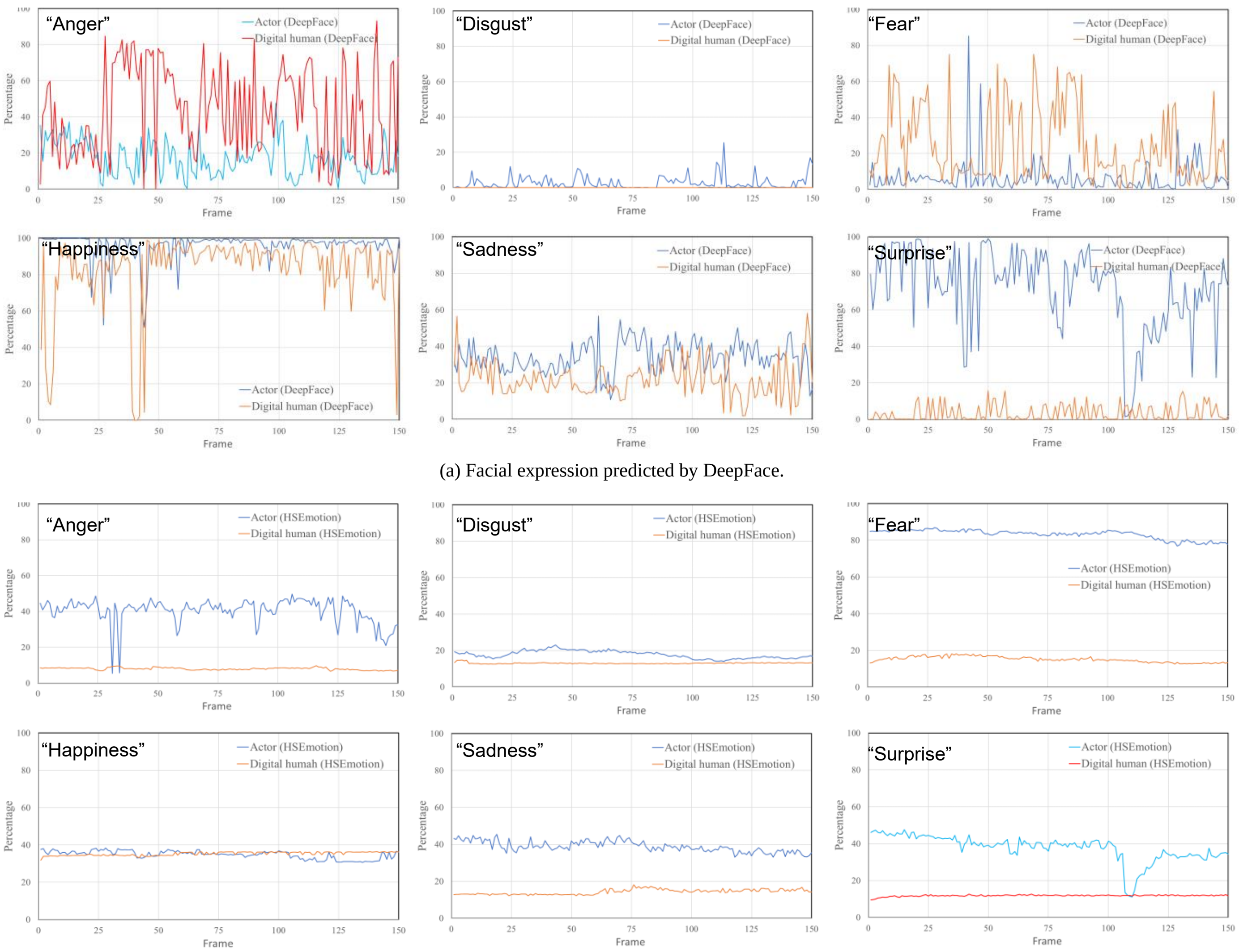
1) Data Taken with LLF

To generate human digital facial expression data, six basic facial expression images were presented in sequence to a 22-year-old Japanese female actor, who was asked to mimic each facial expression for 5 seconds. The facial images were acquired with their features recorded at 30 fps using LLF in order to analyze the changes in facial expressions over time. To minimize the effects of cultural differences and the actor's experience with facial expression, a set of facial images representing Ekman's six basic emotions [10] was presented to the actor. The characteristic of facial muscles for each emotion was described for the actor to mimic appropriately. We collected 150 frames of facial images from a 5-second video of each expression, resulting in 900 frames each of facial images of the actor and the digital human imitating the six basic facial expression.

The actor and the generated digital human facial images for each expression are shown in Figure 2. The appearance of the digital human was made to resemble the actor as close as possible. The generated digital humans generally adequately represent the actor's facial expressions, but the detailed expressions tend to look significantly poor, resulting in a weaker negative facial expression.

2) Data Taken with MACD

The MACD was used to acquire a total of 2,142 face images of 17 Japanese students aged 18-22 from the Faculty of Software and Information Science at Iwate Prefectural University, Japan. These facial images are representing six basic facial expressions plus neutral, captured at 21 different angles. As we focused on evaluating how the facial



(a) Facial expression predicted by DeepFace.

(b) Facial expression predicted by HSEmotion.

Figure 5. Recognition rates for the actor's and its corresponding digital human's facial expressions for each frame.

expressions of each student are represented by the digital human, we did not transfer the facial expressions to a digital human that looked exactly like each student, but instead used one common digital human as the transfer destination. MeFaMo was used to capture facial images and their features.

Figures 3 and 4 show two students expressing anger and disgust, respectively, and the digital human corresponding to each capture direction. The red rectangle in the center indicates the direction of the frontal view (*pitch* = 0, *yaw* = 0). Here, different students' facial expressions are transferred to the same digital human, showing that each student's expression is well represented.

IV. Evaluation of the Quality of Facial Expression

The facial expression data created in Section III is evaluated using the deep learning-based facial expression recognition libraries: DeepFace [13] and HSEmotion [15]. While the architecture of DeepFace simply consists of three convolutional layers and two fully-connected layers, HSEmotion uses eEfficientNet as a backbone, which was fine-tuned on a face identification task using the VGGFace2 dataset. Both libraries perform class classification and use softmax functions in the output layer to normalize the inferred results for each expression from 0 to 1, where the sum of all inferred expression proportions accounts for 1. Hereinafter, this value is called the recognition rate.

DeepFace was trained on the Fec2013 [20] containing 32,298 facial images. In contrast, HSEmotion was trained on the AffectNet [21], a large facial expression dataset containing more than 1,000,000 facial images, and evaluated on datasets such as EmotiW, AFEW (Acted Facial Expression In The Wild), VGAF (Video level Group AFfect) and EngageWild. The trained model "enet_b2_8.pt", which is highly accurate on these datasets, was employed in this study.

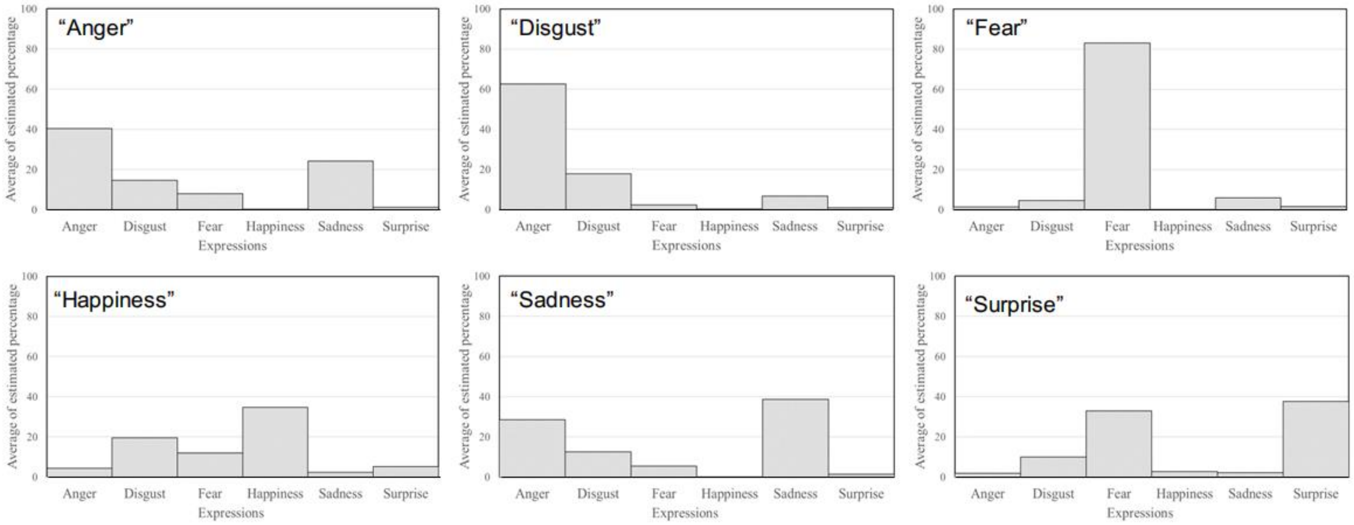


Figure 6. Average recognition rate of the actor's facial expressions by the HSEmotion.

TABLE I. ACTOR'S EXPRESSIONS PREDICTED BY DEEPFACE

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**6**	0	0	0	56	0	0.10
	Disgust	20	0	0	0	128	0	0.00
	Fear	0	0	**2**	7	1	0	0.20
	Happiness	0	0	0	**150**	0	0	1.00
	Sadness	2	0	0	0	**74**	0	0.97
	Surprise	0	0	11	0	0	**135**	0.92
Precision		0.21	NaN	0.15	0.96	0.29	1.00	

TABLE II. DIGITAL HUMAN EXPRESSIONS PREDICTED BY DEEPFACE

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**83**	0	7	0	6	0	0.86
	Disgust	3	0	0	0	19	0	0.00
	Fear	1	0	**38**	0	0	0	0.97
	Happiness	0	0	0	**139**	5	0	0.97
	Sadness	33	0	0	0	**3**	0	0.08
	Surprise	0	0	45	2	0	0	0.00
Precision		0.69	NaN	0.42	0.99	0.09	NaN	

TABLE III. ACTOR'S EXPRESSIONS PREDICTED BY HSEMOTION

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**131**	2	0	0	18	0	0.87
	Disgust	151	0	0	0	0	0	0.00
	Fear	0	0	**151**	0	0	0	1.00
	Happiness	0	0	0	**151**	0	0	1.00
	Sadness	6	0	0	0	**145**	0	0.96
	Surprise	0	0	39	0	0	**106**	0.73
Precision		0.45	0.00	0.79	1.00	0.89	1.00	

TABLE IV. DIGITAL HUMAN EXPRESSIONS PREDICTED BY HSEMOTION

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	0	0	0	0	0	0	NaN
	Disgust	0	0	0	0	0	0	NaN
	Fear	0	0	0	0	0	0	NaN
	Happiness	0	0	0	**150**	0	0	1.00
	Sadness	0	0	0	0	0	0	NaN
	Surprise	0	0	0	0	0	0	NaN
Precision		NaN	NaN	NaN	1.00	NaN	NaN	

A. *LLF-captured Facial Espressions*

Facial expression recognition was performed for each video of facial expression of the actor captured by the LLF and the corresponding digital human. Figure 5 shows the facial expression recognition rates obtained for a 5-second facial expression. The blue and red lines are the results for the actor and the digital human, respectively.

Overall, the results of facial expression recognition for DeepFace were more variable than those for HSEmotion, suggesting that its recognition is unstable. For a more detailed analysis, we investigate results based on their recognition rates using the following criteria.

1) 50% rate for a Correct Recognition

In a given frame, an expression is considered to be correctly judged by the expression recognition library if the recognition rate is 50% or more. According to this criterion, DeepFace was found to recognize "Fear," "Happiness," "Sadness," and "Surprise" of the actor and "Anger," "Fear," "Happiness," and "Sadness" for the digital human from the 5-second video, as shown in Figure 5(a). On the other hand, as shown in Figure 5(b), HSEmotion could only recognize the actor's "Fear."

At first glance, the results for HSEmotion appear to be very poor. However, the results for the actor show that the average recognition rate of its facial expressions is higher than that of the others, with the exception of "Disgust," as shown in Figure 6. The results show that HSEmotion fairly

TABLE V. AVERAGE OF DEEPFACE RECOGNITION RATES FOR 17 SUBJECTS' FACIAL EXPRESSIONS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**18.3%**	0.2%	9.1%	1.4%	44.2%	26.5%	0.1%
	Disgust	18.8%	**0.2%**	9.2%	1.4%	43.1%	27.2%	0.1%
	Fear	19.7%	0.2%	**9.4%**	1.1%	41.0%	28.4%	0.1%
	Happiness	19.9%	0.2%	9.5%	**1.7%**	39.9%	28.5%	0.1%
	Neutral	20.0%	0.2%	9.1%	2.3%	**39.9%**	28.4%	0.2%
	Sadness	19.5%	0.2%	8.3%	2.2%	42.6%	**27.2%**	0.1%
	Surprise	17.6%	0.2%	8.1%	2.3%	45.4%	26.3%	**0.1%**

TABLE VI. AVERAGE OF DEEPFACE RECOGNITION RATES FOR DIGITAL HUMAN FACIAL EXPRESSIONS FROM 17 SUBJECTS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**9.7%**	1.2%	18.0%	13.9%	33.7%	19.1%	4.0%
	Disgust	8.9%	**1.2%**	17.9%	13.8%	33.7%	19.9%	4.1%
	Fear	8.4%	0.8%	**17.9%**	13.3%	34.4%	20.2%	4.3%
	Happiness	8.2%	0.8%	18.3%	**13.7%**	34.2%	19.8%	4.3%
	Neutral	8.4%	1.3%	18.7%	13.9%	**33.6%**	19.4%	4.2%
	Sadness	9.2%	1.4%	19.1%	13.4%	33.0%	**19.4%**	4.1%
	Surprise	9.4%	1.4%	19.6%	13.1%	32.7%	20.1%	**3.1%**

TABLE VII. AVERAGE OF HSEMOTION RECOGNITION RATES OF 17 SUBJECTS' FACIAL EXPRESSIONS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**30.4%**	21.8%	6.1%	2.7%	8.9%	14.9%	2.1%
	Disgust	20.1%	**29.4%**	5.4%	3.9%	8.8%	16.4%	2.3%
	Fear	8.3%	19.2%	**16.5%**	5.6%	15.3%	10.9%	11.3%
	Happiness	4.7%	17.2%	9.4%	**24.2%**	9.3%	5.0%	4.0%
	Neutral	12.5%	14.9%	6.9%	2.4%	**28.2%**	14.3%	3.3%
	Sadness	16.3%	16.5%	6.2%	5.4%	14.1%	**21.2%**	2.5%
	Surprise	4.2%	13.6%	15.6%	7.5%	10.9%	3.7%	**33.8%**

TABLE VIII. AVERAGE OF HSEMOTION RECOGNITION RATES OF DIGITAL HUMAN FACIAL EXPRESSIONS FROM 17 SUBJECTS CAPTURED BY MACD.

		Prediction						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**6.7%**	44.0%	7.1%	1.7%	3.0%	23.4%	2.0%
	Disgust	6.3%	**42.6%**	9.0%	2.6%	2.5%	23.5%	2.8%
	Fear	7.1%	29.5%	**22.5%**	1.2%	2.9%	21.9%	5.4%
	Happiness	5.6%	32.1%	15.2%	**2.4%**	2.7%	25.8%	2.6%
	Neutral	4.0%	34.0%	11.9%	0.7%	**3.4%**	33.2%	2.3%
	Sadness	4.0%	43.5%	8.2%	1.4%	2.5%	**28.3%**	1.9%
	Surprise	7.2%	25.1%	29.3%	1.1%	3.2%	16.7%	**8.4%**

recognized most of the actor's facial expressions. Therefore, we believe that judging the largest recognition rate of each expression as the dominant expression would be preferable, as described in the next sub-section.

2) Highest rate for a Correct Recognition

Facial expression was applied to each frame of the 5-second facial expression video, and the expression with the highest recognition rate was identified as the final recognition result. These results are summarized as confusion matrices in Tables I to IV. The numbers underlined in bold indicate the number of frames in which the actor's facial expression was correctly recognized.

Table I shows that DeepFace has high Precision and Recall for the actor’s “Happiness” and “Surprise.” Recall is high for the actor's “Sadness,” but Precision is low. On the other hand, in digital humans, Table II shows that DeepFace has high Precision and Recall in “Anger” and “Happiness.” In addition, the Recall is high in “Fear.” In terms of high Precision and Recall, we observed that digital humans only accurately represent actors expressing "Happiness."

TABLE IX. RECOGNITION RATES IN DEEPFACE FOR SUBJECTS'FACIAL EXPRESSIONS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev		Yaw							St.Dev
	-20°	-0°	20°			-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	15.5%	21.8%	17.7%	3.2%	Anger	23.9%	12.9%	13.0%	7.2%	13.2%	24.1%	33.9%	9.2%
Disgust	0.4%	0.2%	0.0%	0.2%	Disgust	0.0%	0.3%	0.5%	0.3%	0.3%	0.0%	0.0%	0.2%
Fear	6.7%	9.3%	15.7%	4.6%	Fear	17.0%	10.6%	10.2%	9.0%	10.8%	7.5%	8.7%	3.1%
Happiness	44.6%	43.6%	40.8%	2.0%	Happiness	35.8%	37.3%	27.7%	38.6%	46.8%	59.9%	56.2%	11.7%
Sadness	42.4%	19.3%	18.4%	13.6%	Sadness	26.4%	19.3%	16.1%	29.0%	38.1%	33.5%	24.3%	7.7%
Surprise	7.0%	30.6%	28.8%	13.1%	Surprise	18.0%	24.0%	25.0%	22.0%	19.3%	26.3%	20.3%	3.1%

Table X. RECOGNITION RATES IN DEEPFACE FOR FACIAL EXPRESSIONS OF DIGITAL HUMANS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev		Yaw							St.Dev
	-20°	-0°	20°			-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	6.5%	2.8%	19.7%	8.9%	Anger	10.4%	11.6%	15.6%	8.8%	9.6%	4.5%	7.3%	3.5%
Disgust	1.5%	0.5%	2.3%	0.9%	Disgust	1.0%	5.2%	2.5%	1.2%	0.0%	0.1%	0.0%	1.9%
Fear	18.8%	27.4%	15.3%	6.2%	Fear	30.7%	21.7%	20.3%	25.1%	19.6%	11.6%	14.6%	6.3%
Happiness	11.5%	20.8%	24.5%	6.7%	Happiness	29.3%	25.8%	26.0%	15.4%	26.4%	22.3%	19.0%	4.8%
Sadness	34.4%	9.7%	6.0%	15.5%	Sadness	8.4%	17.1%	25.1%	27.6%	17.4%	12.9%	8.4%	7.6%
Surprise	3.9%	31.5%	27.1%	14.8%	Surprise	14.0%	8.7%	8.7%	23.4%	30.6%	35.6%	24.8%	10.6%

Table XI. RECOGNITION RATES IN HSEMOTION FOR SUBJECTS' FACIAL EXPRESSIONS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev		Yaw							St.Dev
	-20°	-0°	20°			-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	29.8%	33.9%	27.6%	3.2%	Anger	29.5%	31.1%	30.9%	31.0%	29.3%	31.5%	29.7%	0.9%
Disgust	24.6%	33.8%	29.9%	4.6%	Disgust	29.3%	29.0%	29.8%	29.8%	27.4%	30.0%	30.7%	1.1%
Fear	15.2%	16.8%	17.4%	1.2%	Fear	18.1%	16.8%	17.4%	17.5%	16.0%	15.1%	14.4%	1.4%
Happiness	24.0%	24.8%	23.8%	0.6%	Happiness	23.9%	24.3%	23.9%	24.2%	24.6%	24.5%	24.0%	0.3%
Sadness	30.6%	19.8%	13.3%	8.8%	Sadness	22.6%	21.8%	23.0%	21.0%	22.8%	18.5%	18.8%	1.9%
Surprise	29.4%	36.3%	35.7%	3.8%	Surprise	34.3%	35.5%	34.0%	34.9%	30.3%	33.8%	33.5%	1.7%

Table XII. RECOGNITION RATES IN HSEMOTION FOR FACIAL EXPRESSIONS OF DIGITAL HUMANS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev		Yaw							St.Dev
	-20°	-0°	20°			-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	6.2%	6.6%	7.3%	0.6%	Anger	5.0%	4.9%	6.7%	7.7%	6.6%	7.7%	8.4%	1.4%
Disgust	27.3%	43.9%	56.5%	14.6%	Disgust	45.6%	45.7%	48.2%	43.4%	39.7%	37.5%	37.8%	4.2%
Fear	17.2%	24.6%	25.7%	4.6%	Fear	28.0%	27.2%	24.4%	22.5%	21.2%	17.2%	16.8%	4.5%
Happiness	1.9%	2.0%	3.4%	0.8%	Happiness	2.9%	2.4%	2.8%	2.2%	2.7%	2.2%	1.8%	0.4%
Sadness	47.2%	27.7%	9.9%	18.7%	Sadness	30.2%	29.2%	23.6%	22.0%	27.2%	31.6%	34.3%	4.4%
Surprise	7.2%	9.1%	9.0%	1.1%	Surprise	7.3%	8.3%	10.5%	11.5%	6.9%	7.7%	6.6%	1.9%

In contrast to DeepFace, HSEmotion produces high Precision and Recall for the actor's "Anger," "Fear," "Happiness," "Sadness," and "Surprise," as shown in Table III. However, since only "Happiness" is recognized by the digital human (Table IV), we can consider that the digital human only expresses the actor's "Happiness" accurately, as mentioned above. This result is consistent with the findings of our previous study [1].

B. *MACD-captured Facial Expressions*

Tables V to VIII shows the average of the facial expression recognition rates of the 17 subjects and their corresponding digital humans captured by MACD, respectively. The numbers underlined in bold indicate the rate that the expressed facial expression was judged correctly, and shaded areas indicate high rates of incorrect recognition of the expressed facial expression.

The recognition results of DeepFace showed that most facial expressions of the subjects were judged as Neutral with the high rates (Table V). The same trend can be seen in the recognition results of the digital humans for these subjects (Table VI). These results differ significantly from the recognition results of the actor captured by LLF and its corresponding digital human.

On the other hand, HSEmotion fairly recognized the facial expressions of the subjects (Table VII). Only the subjects' "Fear" was relatively misrecognized as "Disgust." However, HSEmotion failed to recognize facial expressions of the digital humans (Table VIII).

To summarize, HSEmotion has more stable recognition rates compared to DeepFace for the same facial expressions of subjects who were captured from different directions. However, when these subjects' facial expressions were transferred to digital humans, both libraries failed to recognize their facial expressions.

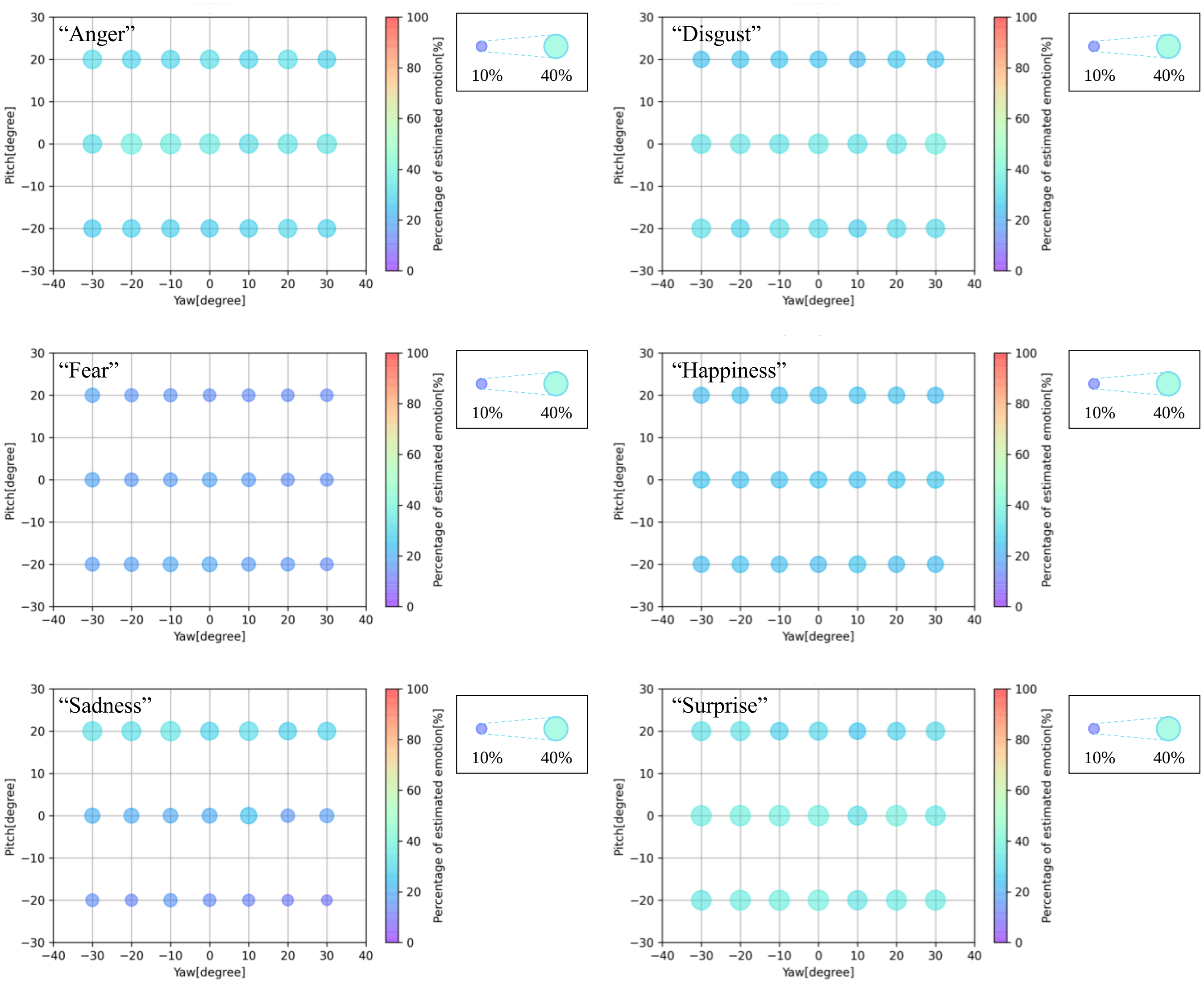


Figure 7. MACD-captured subjects' facial expressions recognized by the HSEmotion.

For a more detailed analysis, we averaged the facial expression recognition rates for three pitch- and six yaw-directions and examined the variation in facial expressions recognized for each direction (Tables IX to XII). Here, we used the Standard Deviation (St.Dev) to represent the variation in the recognition rates. Shaded areas indicate areas where variations in recognition rates differs by 5% or more depending on the camera angles, despite the same facial expression.

The changes of recognition rates in pitch- and yaw-directions are greater for DeepFace. On the other hand, HSEmotion showed less variation in recognition rates for facial expressions captured in different yaw directions. Finally, as shown in Figure 7, the facial expressions of the subjects captured by MACD were recognized by HSEmotion, and the average recognition rate for each angle was visualized. To intuitively see the recognition rate for each angle, the rate is represented by a circle and a color. From this figure, less variation in the recognition rates in the yaw-direction can be observed. These results of HSEmotion benefit from the use of eEfficientNet as its backbone and the large training dataset.

To obtain better facial expression recognition for our experiment, it is necessary to include digital human facial expression images in the training data. However, the limitations of this experiment were due to the fact that the MHC face images could not be used as training data by its agreement. Alternatives to MHC that can be used as training data are needed for further experiments.

V. Conclusion

The widespread use of digital humans has led to a demand to evaluate the facial expressions of them. In this study, we focused on the reality of facial expressions of digital humans and evaluated the results of their facial expression using existing deep learning-based automated facial expression recognitions.

Two strategies were employed in the creation of the dataset to allow for a more detailed analysis. First, we created a dataset of facial expression images by an actor whose facial expressions were clearly expressed. Second, we created a dataset for 17 subjects' facial expressions using multi-angle camera devices. We then created a digital human based on this dataset, and objectively evaluated the reality of their facial expressions using two deep learning-based facial expression recognition libraries.

The evaluation using DeepFace and HSEmotion revealed that the accuracy of facial expression recognition by these libraries was problematic. HSEmotion was effective in evaluating the facial expressions of actual people. The library also absorbed differences in capture direction, providing stable recognition of facial expressions. However, we found that these libraries are not sufficient for evaluating expressions by digital humans, indicating the need for the development of better expression recognition libraries in the future. The terms of use of digital human in this study do not permit training on facial expressions by digital human, which prevented us from creating our own training model based on this dataset.

Future work will include a facial expression recognition library that considers both real people and digital humans. We would like to improve the facial expression recognition model using trainable digital humans to develop a facial expression recognition library that takes both actual people and digital humans into consideration.

REFERENCES

[1] S. Kikuchi, O. D. A. Prima, and H. Ito, “Do Digital Human Facial Expressions Represent Real Human's?,” The Fifteenth International Conference on Advances in Computer-Human Interactions, ACHI2022, pp. 1-5, 2022.

[2] Virtual Shibuya, https://news.kddi.com/kddi/corporate/newsrelease/2020/05/15/4437.html [retrieved: December, 2022]

[3] Juntendo Virtual Hospital, https://jp.newsroom.ibm.com/2022-04-13-Juntendo-Virtual-Hospital [retrieved: December, 2022]

[4] UNEEQ, https://digitalhumans.com [retrieved: October, 2022]

[5] Y. Iwayama, “Real Avatar Production - Raspberry Pi Zero W Based Low-Cost Full Body 3D Scan System Kit for VRM Format,” 10th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, pp. 22-23, 2019.

[6] MediaPipe, https://google.github.io/mediapipe/ [retrieved: December, 2022]

[7] Digital Andy Serkis, https://www.unrealengine.com/en-US/blog/epic-games-and-3lateral-introduce-digital-andy-serkis [retrieved: December, 2022]

[8] Character Creator, https://www.reallusion.com/ [retrieved: December, 2022]

[9] Buddy Builder, https://www.hologress.com/ [retrieved: December, 2022]

[10] P. Ekman and W. V. Friesen, “Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions,” Malor Books, 2003.

[11] T. Kudoh and D. Matsumoto, “The Emotional World of the Japanese - Uncovering the Mysteries of Their Mysterious Culture,” Seishinshobo, 1996.

[12] T. Ohta, M. Tamura, M. Arita, N. Kiso, and Y. Saeki, “Facial Expression Analysis : Comparison with Results of Paul Ekman,” Special Issue for the 10th Anniversary of the Faculty of Nursing), 3(1), pp. 20-24, 2005. (in Japanese)

[13] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1-5, 2020.

[14] A. V. Savchenko, “High-Speed Emotion Recognition Library,” Software Impacts, 14, pp. 1-4, 2022. https://doi.org/10.1016/j.simpa.2022.100433

[15] HSEmotion , https://github.com/HSE-asavchenko/face-emotion-recognition [retrieved: December, 2022]

[16] S. H. Kang and J. H. Watt, “The Impact of Avatar Realism and Anonymity on Effective Communication Via Mobile Devices,” Computers in Human Behavior, 29(3), pp. 1169-1181, 2013.

[17] M. Grewe et al., “Statistical Learning of Facial Expressions Improves Realism of Animated Avatar Faces,” Frontiers in Virtual Reality, 2, pp. 1-13, 2021.

[18] Meta Human Creator, https://www.unrealengine.com/en-US/metahuman-creator [retrieved: December, 2022]

[19] MeFaMo - MediaPipeFaceMocap, https://github.com/JimWest/MeFaMo [retrieved: December, 2022]

[20] Challenges in Representation Learning: Facial Expression Recognition Challenge, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data [retrieved: December, 2022]

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” IEEE Transactions on Affective Computing, 10(1), pp. 18-31, 2017.

Performance Modeling for Call Centers Providing Online Mental Health Support

Tim Rens de Boer
CWI
Amsterdam, Netherlands
Email: trdb@cwi.nl

Saskia Mérelle
113 Suicide Prevention
Amsterdam, Netherlands
Email: s.merelle@113.nl

Sandjai Bhulai
Vrije Universiteit
Amsterdam, Netherlands
Email: s.bhulai@vu.nl

Rob van der Mei
CWI and Vrije Universiteit
Amsterdam, Netherlands
Email: mei@cwi.nl

***Abstract*—Mental health helplines differ from other call centers, such as customer service call centers, in different aspects. Many of the agents handling conversations are volunteers, the conversations can be described as often more complex and/or emotional, and many of the mental health helplines use a triage system. More understanding is needed to improve staffing and/or service of mental health call helplines. Motivated by this, we propose a call center model that includes the specifics of online mental health helplines, including features such as a triage system for chats and the inclusion of service times consisting of warm-up, conversation, wrap-up and cool-down periods. This call center model is then validated using trace-driven simulation based on various months of real-life (anonymous) call and chat data provided by 113 Suicide Prevention. The model is validated by comparing the waiting times found in the data with the waiting times of the simulation. The results show that the model can simulate the waiting-time performance of the helpline accurately. Secondly, we focus on forecasting the number of arriving chats and telephone calls. Our results show that (Seasonal) Autoregressive Integrated Moving Average ((S)ARIMA) models trained on historical data perform better than other models in the case of short-term forecasting (five weeks or less ahead), while using linear regression works best for long-term forecasts (longer than five weeks). We propose a method on how these daily predictions can be quickly altered for hourly predictions, which can then be used together with the understanding of the model to obtain staffing advice.**

Keywords*—*call center models; queueing; mental health; helplines; data analytics; forecasting; staffing

I. Introduction

This paper is an extension of our previous research presented in [1], which was focused on the proposed call center model and validation. The present paper provides a more elaborate discussion on the analysis of real-life data, the validation of the model using trace-driven simulation, the error-term evaluation of demand forecasting, and how the model and forecasting can be used for staffing purposes.

Mental health helplines are helplines that are concerned with helping or assisting help seekers requiring mental health help, instead of physical care such as emergency helplines. There are many forms of mental health, with many countries having one or multiple helplines. Examples of mental health helplines in the Netherlands are 113 Suicide Prevention [2], the helpline for help-seekers with suicidal thoughts, the listen helpline (Dutch: luisterlijn) [3], and the Kindertelefoon [4] for children. Recently, mental-support helplines have received much attention due to the increase in call volumes related to the (partial) lockdown to combat the spreading of corona [5], [6]. This paper focuses on call center modeling for suicide prevention helplines. However, we emphasize that the results can also be used for modeling the waiting-time performance of other mental health helplines.

Suicide is a worldwide health problem. In 2020, on average, five persons died each day by suicide in the Netherlands alone. Suicide is a leading cause of death among adolescents [7]; worldwide, more than 700,000 people die from suicide every year [8]. In many countries, people struggling with suicidal thoughts can contact a helpline to get support to prevent and reduce suicidal thoughts [9]. In the Netherlands, persons with suicidal thoughts can contact 113, either by telephone or by chat [10], and are helped by volunteers and professionals. It is crucial that help seekers are answered swiftly. Therefore, it is important that adequate staffing is present to answer telephone calls and chat requests, minimizing the waiting times and abandonments (i.e., sudden termination of ongoing calls or chats while waiting). In order to calculate proper staffing levels (e.g., [11]), a good understanding of the processes of the call center system for mental health is required.

Mental health helplines differ in various aspects from classical call centers. First, the subjects of conversations are mental health concerns, e.g., loneliness and substance use [12]–[14]. Second, agents often have to handle complex conversations with the patient-in-need, and may themselves require support during or after a difficult conversation [17]. Therefore, an important aspect is that agents often need some time to cool down after emotionally difficult conversations. Third, when a chat or telephone call enters the system, it has to be determined which agent is best capable of handling the call, which results in a warm-up time before the call is taken into service. Lastly, chat conversations that enter the system first go through a triage phase. The inclusion of triage plays an important role, and functions as a filter [18] to chat requests, checking whether these help seekers are at the right helpline or might require emergency care. The triage employee can also estimate the conversation's difficulty to best match an agent with enough experience to handle the chat. So, the model for such a mental health helpline needs to meet the following requirements: (1) the possibility of abandonments, (2) a service time consisting of multiple phases, (3) a warm-up and cool-down time, (4) inclusion of chats and telephone calls, and (5) triage.

Classic call center models were considered for modeling this helpline, see [19] for an excellent overview of queueing models for call centers. The Erlang-C model is proven to be inaccurate when modeling call centers with abandonments [20]. The Erlang-A model includes abandonments [21], but does not include multiple skills, triage, and warm-up times. Multi-skill call center models, such as an N-configuration [19], were also

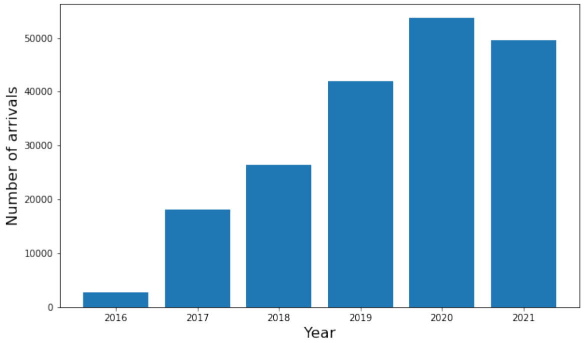


Figure. 1. The number of phone arrivals per year.

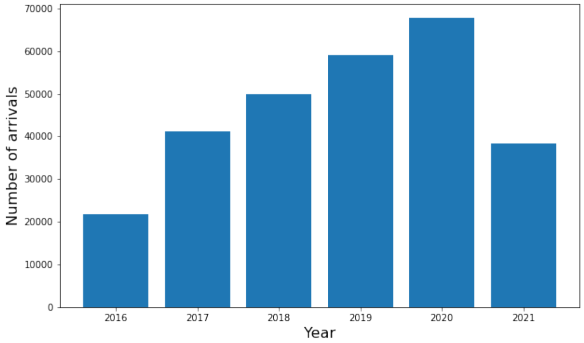


Figure. 2. The number of chat arrivals per year.

considered. The N-configuration model includes two types of arrivals, which in this case would be telephone and chat, and assumes that new chat arrivals can be picked up before triage. However, in the mental health helpline of [2], it is crucial that chats first go through triage. The importance of modeling triage is already shown in the emergency domain [23], [24]. Therefore, in this paper, we modify the N-configuration model in such a way that it includes the specifics of mental health helplines.

This paper introduces a new queuing model to include the possibility of triage. The model is built based on anonymized real-life call and chat data, made available by [2]. The records used cannot be traced back to individual callers, as only timestamps and durations are used. The anonymous data is also used to validate the model and determine different patterns that can be helpful for forecasting demand. Lastly, a method for determining staffing levels is explained for this specific model.

This paper is organized as follows. Section II discussed related work done on either call center models or suicide prevention helplines. Section III describes different patterns found in the data. In Section IV, the proposed model is described. In Section V, the model is validated using a combination of data and trace-driven simulation. Next, Section VI explains how the demand call volumes can be predicted based on historical data. Section VII describes how the model and the forecast can be used to construct staffing advice. Finally, in Section VIII, the conclusion and discussion are given.

II. Related work

Call centers and/or helplines have been researched in different fields of science. For this paper, we will mention some of the research done on the modeling of call centers and research done on specifically suicide prevention helplines. Various research has been done on modeling call centers, Garnett et al. [21] have shown how and why Erlang-B and Erlang-C lack the feature of impatient customers. They introduce an Erlang-B model, where customers abandon the system after not being answered after their impatience has run out. Here, impatience is drawn from an exponential distribution. This research, however, lacks the inclusion of multi-skill. Gans et al. [19] provide an overview of different types of skill routings, where the skill may be based on training, compensation, or time restrictions. Forecasting of arrivals has also been researched, for example, Gijo et al. [22] have shown how (S)ARIMA models can be used in forecasting the call volume. Research done specifically at mental health helplines is often limited to conversation topics and different types of callers. Salmi et al. [14] show the change in conversation topics during COVID-19. While Grigorash et al. [15] and O'Neill et al. [16] have both studied different caller types and how these caller types can be predicted. This paper aims to fill the gap between research done on standard call centers and that done on mental health helplines, specifically by applying modeling and forecasting, that are normally seen in standard call centers, to mental health helplines.

III. Data Analysis

The data for this research is provided by 113 Suicide Prevention [2], referred to as '113'. Its mission is to prevent suicides and break the taboo surrounding suicide. Help seekers struggling with suicidal thoughts can contact 113 24/7 by either chat or telephone. These help seekers are then helped by counselors consisting of volunteers and paid employees. Apart from that, 113 also provides online therapy, self-tests, and self-help courses [25].

Call data is provided over the period 2016-2021 and consists of around 250,000 chats and 175,000 telephone calls. The distribution of the arrivals over the years can be seen in Figures 1 and 2 for phone and chat, respectively. It should be noted that data for 2016 and 2021 were incomplete at the time, explaining the low number of arrivals in 2016 and 2021. However, omitting 2016 and 2021, it can still be seen that there is an increasing trend present in the number of arrivals of phone calls as well as chats. The number of telephone arrivals increased from almost 20,000 in 2017 to more than 50,000 in 2020, and chat arrivals increased from close to 40,000 in 2017 to almost 70,000 in 2020. Each call or chat has the following elements: (1) the arrival time, (2) the time entering the queue, (3) the time of acceptance by an agent, (4) the disconnection time, and (5) the completion time, all elements are denoted using a combination of date and time, using the

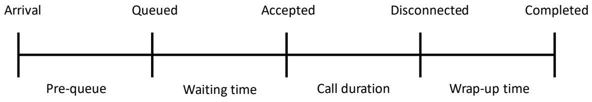


Figure. 3. Timeline of a call from the perspective of the caller.

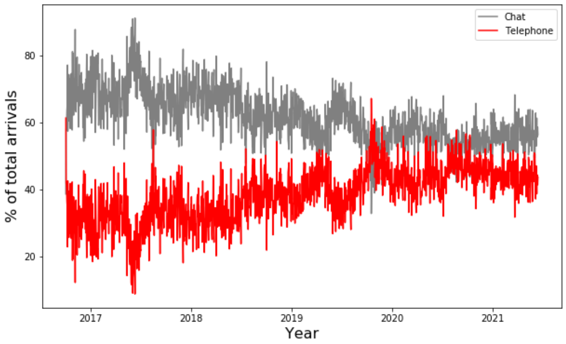


Figure. 4. Decomposition of incoming calls and chats.

following formats dd-mm-yyyy and hh:mm:ss for date and time, respectively. Each call also has a contact id and an *initial* contact id, which may be different if a call is a forwarded telephone call or chat.

Different features were added to the data to obtain more understanding of the different durations found in the data. The following durations were calculated: *pre-queue duration*, *waiting time*, *call duration* and *wrap-up time*. The *pre-queue duration* is the time between the arrival of a call and the time entering the queue, and is the time spent by the caller in a menu. This phase does not require resources from the helpline and, therefore, falls outside the scope of this paper. The *waiting time* is defined as the time between entering the queue and the time that an agent accepts the call. The *call duration* where the agent and the help seeker are actually connected is defined as the time between acceptance of a call and the disconnection time. Finally, after each call, the agent has to fill in a wrap-up form, which is the time between disconnection and completion. A visual representation of this timeline is given in Figure 3.

Recall that there are two options for help seekers to contact the helpline: via *telephone* or *chat*. These two contact types are mostly handled by the same type of agents, but differ in some important aspects. Traditionally, there used to be more chats than telephone calls. However, the difference has diminished in recent years, and the numbers are comparable. This can be seen in Figure 4, which shows that in 2017 and 2018, most of the arrivals were chats. In 2019, the difference between chat and phone diminished, while since 2020, there are still more chats, but the difference is not as large as before. However, the chat and telephone calls do follow different patterns, which are shown in Figures 7 and 8. Figures 5 and 6 show the weekly patterns; in both cases, the weekends see a lower number of calls. However, for chats, we see a clear dip on Saturday and a small increase on Sunday. We also observe that most telephone calls arrive between 12:00 and 20:00, while chats have a clear peak at 20:00. Both telephone and chat call arrivals show a decrease during the night and early morning until around 5:00 in the morning.

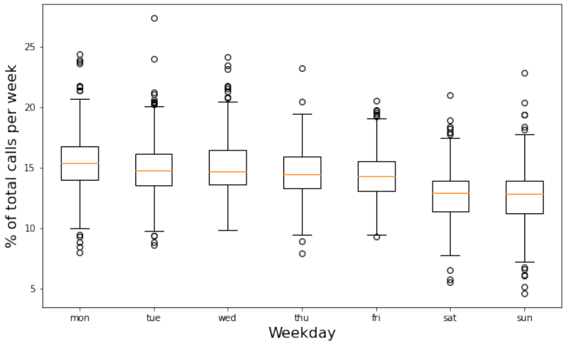


Figure. 5. Weekly pattern of incoming telephone calls.

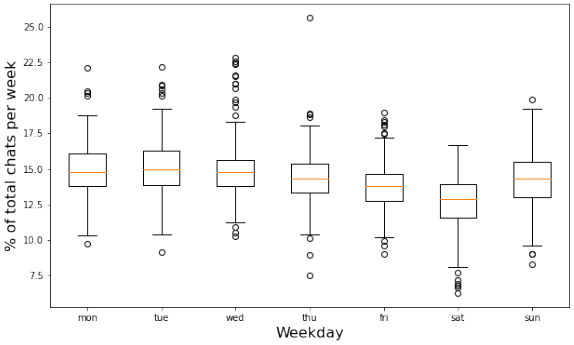


Figure. 6. Weekly pattern of incoming chats.

Incoming chats are first handled by triage, this is needed to filter chats that are at the wrong helpline. Only part of the incoming chats after triage gets sent to another agent. The percentage of chats that an agent handles after triage is around 50% during day shifts, but this differs over the day. During the night shifts, fewer chats are forwarded to an agent. The different nature of night conversations may cause this. The triage plays an important role in filtering chats, as seen by the percentage of chats that get through triage. Chats are filtered out due to various reasons. For example, the chatter may be at the wrong helpline and or is identified as a prank chatter [26].

In the data, we distinguish three service time distributions: for the duration of (1) telephone calls, (2) chats *during* triage, and (3) chats *after* triage. The empirical distribution of phone call durations can be found in Figure 9. On average, a telephone call takes around 18 minutes, with the longest phone call in the data taking multiple hours. The duration of chats in triage can be seen in Figure 10, and the duration of chat conversations after triage can be seen in Figure 11. As can be seen, the chats that have gone through triage tend to take much longer than the chats during triage: on average, chats in triage take 19 minutes versus 34 minutes after triage. This is

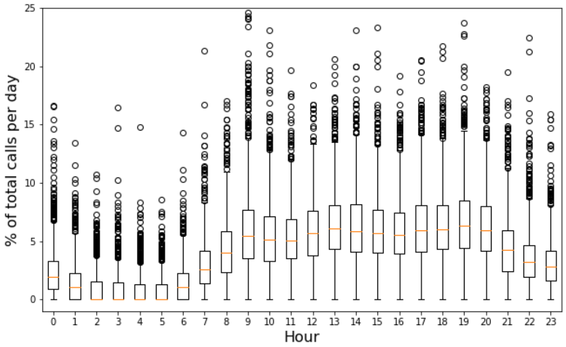


Figure. 7. Daily pattern of incoming telephone calls.

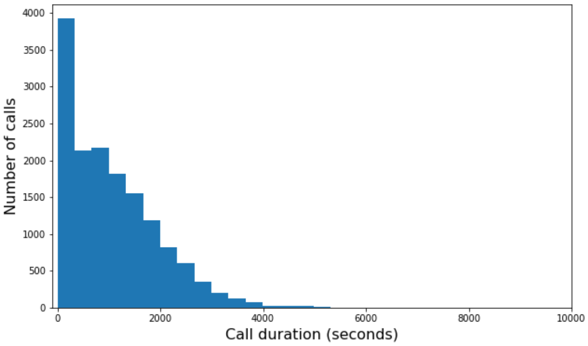


Figure. 9. Histogram of telephone call durations.

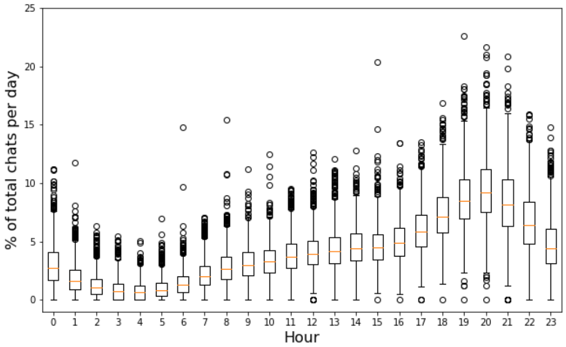


Figure. 8. Daily pattern of incoming chats.

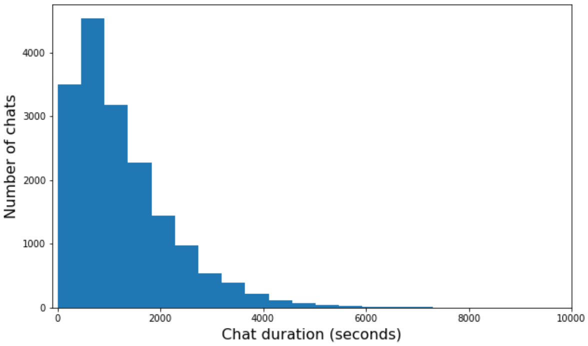


Figure. 10. Histogram of chat triage durations.

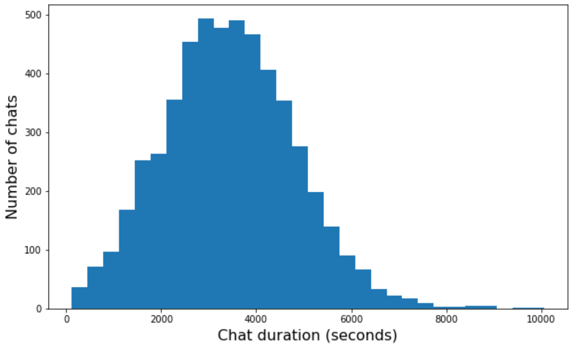


Figure. 11. Histogram of durations of chat conversations after triage.

also clearly visible in the histograms with durations of chats in triage having a clear peak at around 400 seconds, while chats after triage have a peak in durations at 4000 seconds.

Remark: The effect of the experience of agents
Mental health helplines often use a mix of paid professionals and volunteers, where an agent's responsibilities depend on experience and other factors. The assumption is that experienced agents can handle more difficult conversations alone, while inexperienced agents might require assistance when handling a more complex conversation (this could be either a phone or chat conversation), resulting in a longer service time. Analysis of our data pointed out that there seems to be no significant difference in service time distribution between different volunteers and professionals. There might be a difference, but this could be obscured due to many other factors, such as experienced agents handling the more complex conversations and assisting or coaching the less experienced agents. Based on this observation, the call duration distributions are assumed to be the same for all agent experience levels.

IV. Model description

For the modeling step, we distinguish between two different types of calls, namely: (1) *chat calls*, and (2) *telephone calls*, which are handled differently. The incoming chat calls are first handled by a triage system. The triage system consists of c_{triage} triage agents. Each triage agent can handle n_{triage} chats simultaneously without perceivable slow-down. Together, there are $c_{\text{triage}} * n_{\text{triage}}$ 'triage slots' for incoming chats. Arriving chats enter service at triage immediately if a triage slot is available. If no triage slot is available, then chats enter an infinitely sized queue and are helped first-come-first-served (FCFS) or abandon the queue if the waiting time is longer than their impatience. The service time of a chat call in triage, denoted B_{chat}, is the convolution of four random variables:

$$B_{\text{chat}} = B_{\text{warm-up}} + B_{\text{conversation}} + B_{\text{wrap\_up}} + B_{\text{cool-down}}.$$

All variables are drawn from some probability distribution, which can be estimated from the data. A visual representation of the service time can be seen in Figure 12.

A key aspect of the model is the inclusion of *impatience*, i.e., the maximum amount of time that a help seeker is willing to wait before he abandons the system. The impatience of a help seeker who enters the system via a chat is modeled as an independent sample from some probability distribution with mean $\mu_{\text{chat-impatience}}$. After service completion at the triage center, a chat is either sent through to the helpline (HL) for assistance (with probability $p_{\text{sent-through}}$), or the chat leaves the system. Note that during the warm-up period $B_{\text{warm-up}}$, the agent is busy, but the help seeker is not yet answered. Therefore, help seekers may abandon the queue during that period.

Different from chat calls, incoming telephone calls do *not* go through a triage phase and arrive directly at the HL. This is the core part of the system where most of the service processing occurs. The HL is equipped with c_{HL} agents, each of which can handle both chat calls and telephone calls, not more than one call at a time. When a telephone call finds an HL-agent available, he enters service immediately. If the telephone call arrives and no agent is available, the call enters an infinite-sized queue that is handled on an FCFS basis. Here, phone calls are able to abandon the queue if their waiting time is longer than their impatience.

The HL processes both telephone calls and forwarded chat calls (i.e., those that have passed through the triage phase). Here, *chat calls have non-preemptive priority over telephone calls*. Thus, when an HL-agent becomes idle, he first checks whether there is a chat call pending (while keeping a triage slot occupied), and if so, starts to service the longest-waiting chat call. If no chat call is pending, the agent checks if telephone calls are pending.

Similar to the modeling of chat sessions in triage, the duration of phone calls and chats after triage both consists of four subsequent independent phases: (1) warm-up, (2) conversation, (3) wrap-up, and (4) cool-down, where each phase has its own probability distribution differently for phone and chat after triage. The impatience of help seekers via telephone is modeled as a sample from some probability distribution that can be obtained from the data. Calls abandon the queue if their waiting time exceeds the impatience. When service is completed, the calls exit the system. Note that agents are busy during the warm-up period, similar to abandonments at the triage, but the help seeker does not perceive this and can, therefore, still abandon the queue during the warm-up phase. See Figure 13 for an illustration of the model.

V. Model validation

The model described in the previous section has to be validated to confirm our modeling choices. The validation was done using trace-driven simulation. For the simulation of this model, the following values are needed for each arrival: (1) an arrival time, (2) service duration (consisting of warm-up, conversation, wrap-up and cooldown), (3) impatience, and (4) call type (chat or telephone) and if the call type is chat information about triage (duration and if the chat is at the right helpline) are also needed. Trace-driven simulation uses values obtained exactly from the data, thereby giving a precise comparison between reality and simulation. The data contained some missing values, and these need to be filled in before simulating, namely: (a) the conversation and wrap-up duration of calls that were unanswered, (b) impatience of help seekers, and (c) warm-up and cool-down durations.

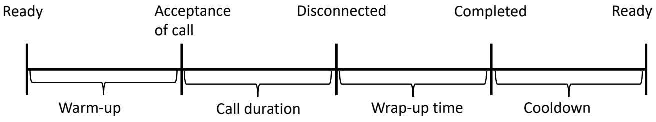


Figure. 12. Timeline of call from the perspective of the agent.

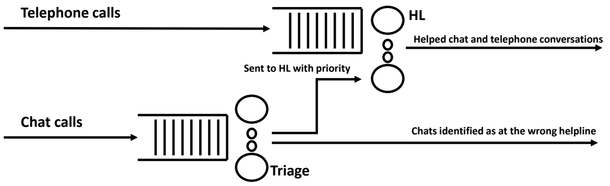

Telephone calls
HL
Helped chat and telephone conversations
Sent to HL with priority
Chat calls
Chats identified as at the wrong helpline
Triage

Figure. 13. Illustration of the model.

The missing values of conversation and wrap-up durations were filled in using hot-deck-imputation [28]. This method samples from the known values to fill in the missing values. The distinction was made between the different types of conversations: telephone, chats during triage, and chats after triage. The impatience of help seekers are mainly unknown due to the limited availability. Only a small percentage of telephone calls were unanswered, and for chats, even fewer impatience data was available. Warm-up and cool-down durations were not present in the data. Therefore, the missing values of impatience, warm-up, and cool-down periods were all drawn from exponential distributions.

The parameters were estimated using expert opinions from paid professionals and volunteers. Impatience of chat conversation is determined by the sum of a constant 300 seconds and a duration drawn from an exponential distribution with a mean of 300 seconds. The impatience of a telephone caller is drawn from an exponential distribution with a mean of 240 seconds. The warm-up for both chat and telephone is drawn from exponential distributions with means of 60 and 45 seconds for telephone and chat, respectively. Lastly, the cool-down durations of chat and telephone are both drawn from an exponential distribution with a mean of 120 seconds.

Moreover, based on current practice at 113, for the experiments, we assume that $n_{\text{triage}} = 5$. Thus, each triage agent can handle a maximum of five triage chats simultaneously. Further, the simulations are trace-driven, and follow the realizations of the time variations of (1) the arrival processes for chat and

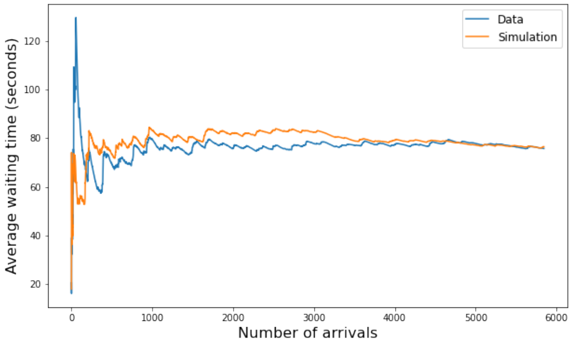


Figure. 14. The average waiting time of telephone calls in the simulation and the data of August 2021.

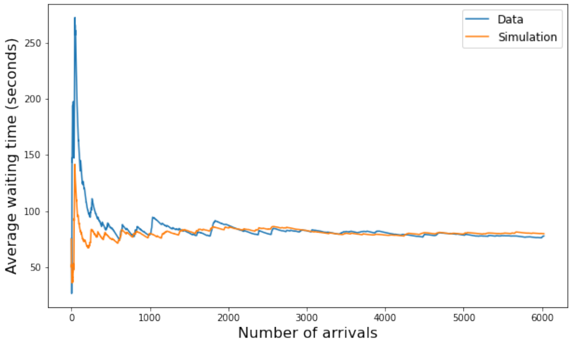


Figure. 15. The average waiting time of chat calls in the simulation and the data of August 2021.

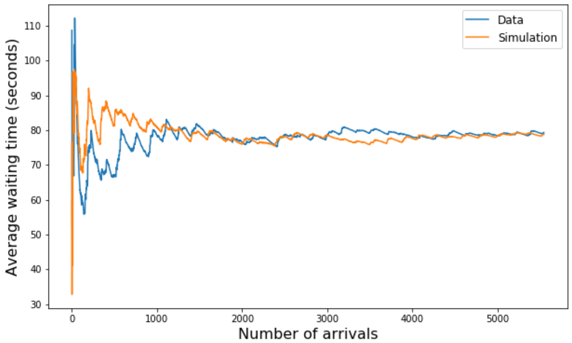


Figure. 16. The average waiting time of telephone calls in the simulation and the data of September 2021.

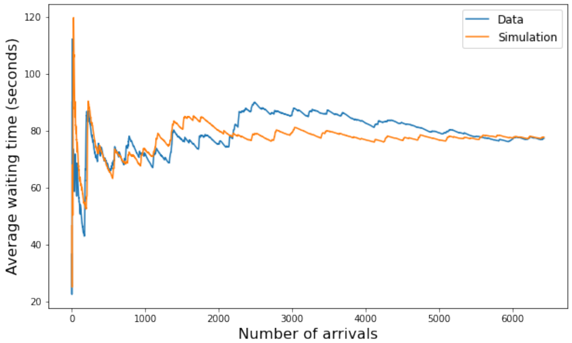


Figure. 17. The average waiting time of chat calls in the simulation and the data of September 2021.

telephone, (2) the number of triage- and HL-agents, and (3) the fraction of triage call that is forwarded to the HL system.

The simulation ran for almost 12,000 arrivals consisting of around 6,500 chat arrivals and 5,500 phone calls during a period of 1 month. The average waiting time of the simulation and the data were then compared. Figure 14 shows both the simulated and realized average waiting times for telephone calls as a function of the cumulative number of call arrivals. The results show that for a small number of calls, the average waiting time is rather sensitive to outliers but quickly stabilizes over time. The average waiting time of the data and simulation both converge to the same value, around 80 seconds. This validation experiment is repeated for chat call arrivals, which show similar convergence, see Figure 15. This process was repeated over different months and showed similar convergences; an example can be seen in Figures 16 and 17, which shows the experiment repeated for September.

In summary, these validation results show that the model works well in predicting waiting times and confirms the modeling choices made in Section IV.

VI. Demand forecasting

Demand forecasting concerns itself with predicting the call volumes for both telephone and chat calls. The performance of the forecasts is dependent on several choices: the time window and the aggregation level. The time window concerns itself with how long ahead the forecast is. Forecasting daily volumes for tomorrow is often easier than the daily volume over 8 weeks. For this paper, we chose to predict 1 day ahead until 8 weeks ahead, since the schedule of the agents is made 8 weeks ahead but can be adapted over time. The aggregation level concerns itself with the kind of volumes to be forecasted. Hourly volumes are harder to predict than monthly volumes, but are less useful for scheduling. In this paper, we chose to predict daily volumes. However, later in this section, it is explained how these daily volumes can be adapted to hourly volume predictions. The choice was made to forecast chat and telephone arrivals independently due to the difference in arrival processes (also explained in Section III) and the difference in handling.

The following forecasting models were considered:(1) Long Short-Term Memory (LSTM), (S)ARIMA [29], linear regression [30], and different baseline models. The parameters of (S)ARIMA models are chosen using auto-ARIMA [31], which are $(5, 1, 1)$ for ARIMA and $(1, 1, 1)(0, 1, [1, 2], 7)$ for SARIMA. As baseline, the prediction of day i is the volume measured on day $i-7$ (baseline model 1) and $i-58$ (baseline model 2). The following aspects were seen as important: *trend* and *seasonality*. The provided data shows that there is an increasing trend present and that weekly cycles seem to be predominant. Therefore, it is chosen to focus on forecasting

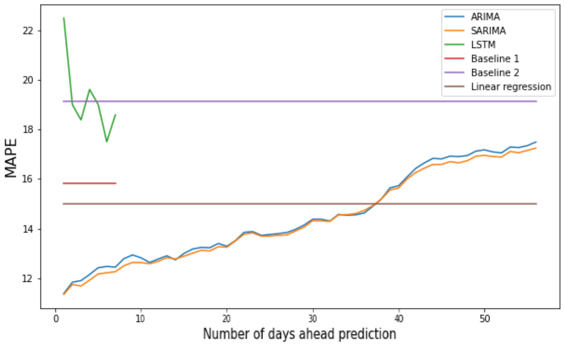


Figure. 18. The MAPE error when forecasting telephone.

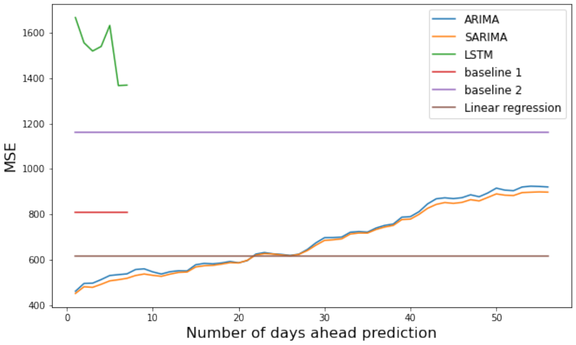


Figure. 20. The MSE error when forecasting telephone.

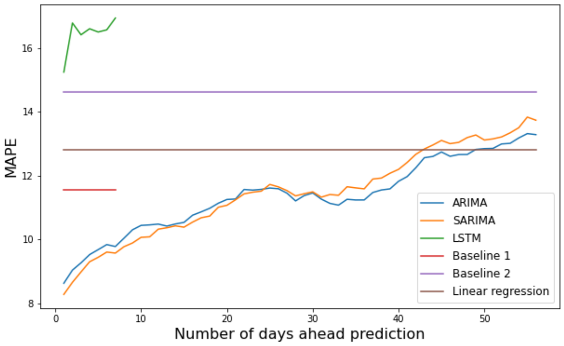


Figure. 19. The MAPE error when forecasting chat.

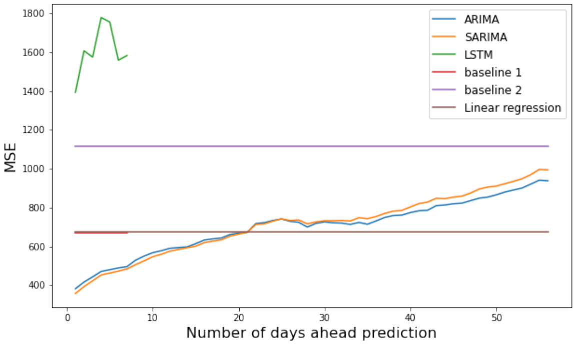


Figure. 21. The MSE error when forecasting chat.

using historical data on call volumes. The models will be evaluated using the Mean Absolute Percentage Error (MAPE), the Mean Squared Error (MSE), and the Mean Absolute Error (MAE). The MAPE is calculated using the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|\frac{F_i - A_i}{A_i}| * 100\%.$$

Here n is the number of forecasts, F_i is the forecast of day i, and A_i is the actually recorded number of arrivals on day i. The found MAPE results can be seen in Figures 18 and 19. We find that (S)ARIMA models perform best (in terms of the MAPE), especially when forecasting for five weeks (or six in the case of chats) ahead or less. For longer forecasting windows, it turns out that a simple linear regression model might provide more accurate forecasts in the case of telephone arrivals. However, both (S)ARIMA and linear regression models have a MAPE that is lower than the MAPE of the baseline models. The LSTM model has the highest error term in this situation.

Next, the models are evaluated using MSE, which gives more weight to large prediction errors. The MSE is calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2.$$

The found MSE values can be found in Figures 20 and 21. The results show that in terms of MSE, (S)ARIMA models have the lowest error. However, for forecasting more than three weeks ahead, the linear regression seems to have a lower error term. This differs from the finding when comparing the MAPE error terms, meaning that the (S)ARIMA models likely have more large prediction error than the linear regression.

Lastly, for a complete comparison, the models are also evaluated using MAE, which is calculated by the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|.$$

The found MAE values can be found in Figures 22 and 23. The graphs show similar findings when evaluating based on MAPE, in the telephone calls as well as chat, (S)ARIMA models perform best when forecasting for short-term windows, and linear regression for longer-term forecasts. The MAE that again forecasting chats is more accurate, and the window when (S)ARIMA performs best is longer (around 40 days for chats versus around 30 days for telephone).

Combining the evaluation of the three different error terms, we can come to the following findings:

1) All error terms agree that for short-term forecasting (S)ARIMA models perform best, long-term forecasting

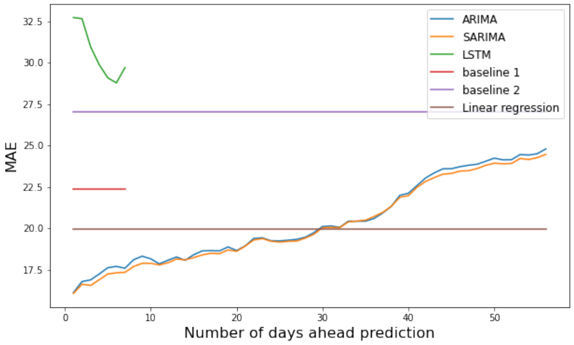


Figure. 22. The MAE error when forecasting telephone.

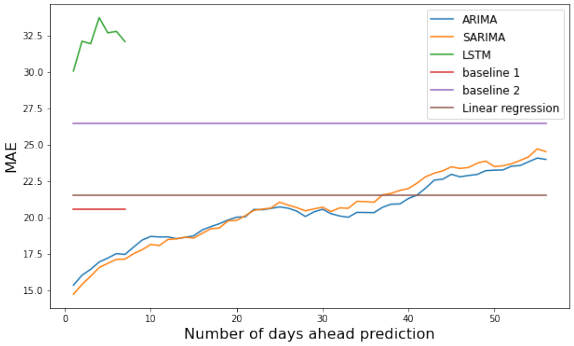


Figure. 23. The MAE error when forecasting chat.

can best be done using linear regression, and LSTM seems to perform the worst.

2) The different error terms differ on the time window of when (S)ARIMA model performs the best. The MSE graph of phone forecasting shows a shorter time window in which (S)ARIMA performs best when compared to that of the MAPE graph. This tells us that it is likely that phone forecasting has more large errors influencing the MSE. A similar picture can be seen for chat forecasting.
3) All evaluations agree that the LSTM model performs the worst of all evaluated models at the moment.

The performance of (S)ARIMA models can be attributed to the flexibility of the models. However, it could be the case that with more time and optimization, the LSTM will perform better. However, it is questionable if the model would perform better than the (S)ARIMA models.

Hourly predictions

The results above concern the prediction of daily volumes. Call centers, however, often use hourly volumes for staffing purposes. Predicting hourly volumes can be done in several ways, namely forecasting hourly volumes directly or indirectly by using the predicted daily volumes. Forecasting hourly volumes directly introduces more uncertainty in the predictions. Therefore, it was chosen to forecast using the daily volumes. The call volume of hour h on day i is calculated as follows:

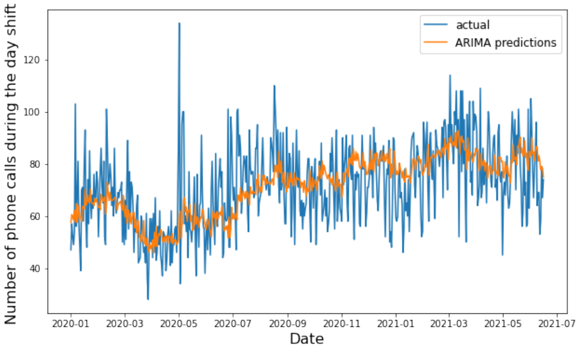


Figure. 24. Comparison between the actual and predicted number of phone calls during the day shift (from 8.00 until 16.00).

$$F_{i,h} = F_i \times P_h.$$

Here P_h is the mean percentage of arrivals on a day that arrives in hour h. This method can be expanded to also include smaller or larger intervals, such as shifts. An example of this can be seen in Figure 24, again the prediction is able to follow the actual number of arrivals.

VII. Staffing

This section will briefly describe how the model of the helpline can be used together with demand forecasting to construct staffing advice. Staffing concerns itself with the required number of agents per time interval and is, therefore, different from a work schedule or planning of agents, for a schedule often staffing advice is first needed. One of the most well-known and often used rules is the square-root staffing rule [32]. This is a formula used to calculate s, the staffing level, and can be rounded up to an integer value. The formula is given by:

$$s = \rho + \beta\sqrt{\rho},$$

where ρ is the offered load calculated by multiplying the arrival rate with the mean service time, β is a parameter reflecting the service level, a higher beta corresponds to a higher quality of service.

The given formula is based on M/M/C queues and can be adapted to other queues. The model described in Section IV needs two different staffing levels at Triage and at HL. The staffing at Triage receives only chat arrivals, ρ_{Triage} is, therefore, calculated by multiplying the chat arrivals with the average service time of chats in Triage. Since staffing is done ahead of time, the chat arrivals are predicted chat arrivals. Together with the fact that agents at Triage can handle five conversations simultaneously, the square-root staffing rule therefore becomes:

$$s_{Triage} = \frac{\rho_{Triage} + \beta \times \sqrt{\rho_{Triage}}}{5}.$$

The staffing at HL can be determined using the square-root staffing rule together with the facts that each agent can handle one conversation and agents handle chats after Triage and new phone calls. The formula then becomes:

$$s = \rho_{HL} + \beta\sqrt{\rho_{HL}}.$$

Here, ρ_{HL} is calculated by summing the chats after Triage multiplied by the average chat service duration, and the phone arrivals with average phone service duration, where again the arrivals are done by forecasting.

VIII. Conclusion

This paper on call center modeling for mental helplines has several contributions. The first contribution is a new call center model for mental helplines. The modeling is based on data and the experience of agents, the model is then validated based on data from [2], the suicide prevention helpline in the Netherlands. The validation is done using trace-driven simulation, and the results show that the model is able to accurately predict waiting-time performance for telephone and chat arrivals at the call center. We emphasize that the model is also applicable to other mental helplines with triage and complex conversations requiring warm-up and cool-down periods, possibly with minor modifications.

The warm-up and cool-down durations are estimated mostly based on experience from agents. There is difficulty in accurately measuring these durations. For future research, these durations could be based on data and possibly correlated to the duration of the call and wrap-up. This could be done by measuring when agents log off and log back on.

In the model, we assume that $n_{triage} = 5$, meaning each agent can handle a maximum of five triage chats simultaneously. Here, we assume that there is no slow down in service time when an agent handles a triage chat more, but is unable to handle a sixth chat. For future research this assumption could be further explored, will the quality decrease and service time increase when setting n_{triage} higher.

The second contribution of this research is centered around demand forecasting. Various models were tested for forecasting the number of arrivals per day and evaluated based on MAPE, MSE, and MAE. All evaluations agreed that for short-term forecasting (S)ARIMA models perform best and linear regression in the case of long-term forecasting. However, the tipping point at which linear regression performs better than (S)ARIMA differs per evaluation. It is also discussed how these daily volume predictions can be changed into hourly or shift volumes.

Lastly, the third contribution is on how to combine the previous two contributions (modeling and forecasting) for staffing purposes. We explain how the model and forecast can be used to adapt the square-root staffing rule, which in turn can be used to generate staffing advice. Regarding the real-life applicability of this paper, the staffing advice and forecasts obtained can easily be combined into a user-friendly dashboard, which can easily be used by the planning department to possibly improve staffing.

It is also important to note that while the staffing advice is important for the performance of the helpline, there are also other ways to reduce the waiting time or increase the percentage of answered telephone and chat calls. Some of these methods are already used by 113 at the moment, such as staff in a different time zone, shortening the menu a new help-seeker has to listen, aiming for shorter conversations and staff extra during the switch moments of different shifts and breaks. For example, 113 Suicide Prevention was able to decrease the waiting time in the night by using call center agents working in a different time zone (Suriname in this case). Another method is shortening the time until a help-seeker enters the queue, by shortening the menu the help-seeker has to listen and fill in, resulting in fewer help-seekers abandoning before entering the queue. Lastly, by aiming for shorter conversations agents are able to handle telephone and chat calls faster, and therefore the overall service time decreases resulting in lower waiting times with the same number of agents or requiring less agents for the same performance.

This paper aims to provide a complete view, but some aspects require follow-up research. For example, the level of experience of volunteers and paid professionals might affect call durations. Preliminary data analysis into this topic shows no, or at best limited, correlation, but further investigation is needed.

Acknowledgment

The authors would like to thank Robin Costers, Sophie de Vries Robles, Menno Zwart, and Kim Setkowski from 113 Suicide Prevention for their great support during this project.

References

[1] T.R. de Boer, S. Mérelle, S. Bhulai and R. D. van der Mei, “A Call Center Model for Online Mental Health Support,” in PREDICTIONS SOLUTIONS 2022, International Conference on Prediction Solutions for Technical and Societal Systems, 2022, pp. 1-6.

[2] “About us 113 Suicide Prevention.” (”Over ons—113 zelfmoordpreventie.”) [Online]. Available: https://www.113.nl/over-113/over-ons (accessed Jun. 24, 2022)

[3] “The listen helpline.” (“De luisterlijn.”) [Online]. Available: https://www.deluisterlijn.nl/ (accessed Jun. 24, 2022)

[4] “The helpline for children.” (“Kindertelefoon.”) [Online]. Available: https://www.kindertelefoon.nl/ (accessed Jun. 24, 2022)

[5] J. Scerri, A. Sammut, S. Cilia Vincenti, P. Grech, M. Galea, C. Scerri, D. Calleja Bitar, and S. Dimech Sant, “Reaching out for help: Calls to a mental health helpline prior to and during the covid-19 pandemic,” International Journal of Environmental Research and Public Health, vol. 18, no. 9, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/9/4505

[6] M. Brülhart, V. Klotzbüucher, R. Lalive, and S. K. Reich, “Mental health concerns during the covid-19 pandemic as revealed by helpline calls,” Nature, vol. 600, no. 7887, pp. 121–126, 2021.

[7] J. Hoogenboezem and T. Traag, “Zelfdoding in Nederland: Een overzicht vanaf 1950,” Aug 2021. [Online]. Available: https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/zelfdoding-in-nederland-een-overzicht-vanaf-1950?onepage=true

[8] WHO, “Suicide.” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[9] M. S. Gould, J. Kalafat, J. L. HarrisMunfakh, and M. Kleinman, "An evaluation of crisis hotline outcomes. part 2: Suicidal callers," Suicide and Life Threatening Behavior, vol. 37, no. 3, pp. 338–352, 2007.

[10] J. K. Mokkenstorm et al., "Evaluation of the 113online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines," Suicide and Life-Threatening Behavior, vol. 47, no. 3, pp. 282–296, 2017.

[11] N. Izady and D. Worthington, "Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments," European Journal of Operational Research, vol. 219, no. 3, pp. 531–540, 2012.

[12] M. E. Pratt, "The future of volunteers in crisis hotline work," Ph.D. dissertation, University of Pittsburgh, 2013.

[13] R. C. W. J. Willems, C. H. C. Drossaert, P. Vuijk, and E. T. Bohlmeijer, "Mental wellbeing in crisis line volunteers: understanding emotional impact of the work, challenges and resources. a qualitative study," International Journal of Qualitative Studies on Health and Well-being, vol. 16, no. 1, 2021, pMID: 34694979. [Online]. Available: https://doi.org/10.1080/17482631.2021.1986920

[14] S. Salmi, S. Mérelle, R. Gilissen, R. D. van der Mei, and S. Bhulai, "Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: in-depth analysis using encoder representations from transformers," BMC public health, vol. 22, no. 1, pp. 1–10, 2022.

[15] A. Grigorash, S. O'Neill, R. Bond, C. Ramsey, C. Armour, M. D. Mulvenna et al., "Predicting caller type from a mental health and well-being helpline: analysis of call log data," JMIR Mental Health, vol. 5, no. 2, p. e9946, 2018.

[16] S. O'Neill, R. R. Bond, A. Grigorash, C. Ramsey, C. Armour, and M. D. Mulvenna, "Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline," Health informatics journal, vol. 25, no. 4, pp. 1722–1738, 2019.

[17] F. Sundram, T. Corattur, C. Dong, and K. Zhong, "Motivations, expectations and experiences in being a mental health helplines volunteer," International journal of environmental research and public health, vol. 15, no. 10, p. 2123, 2018.

[18] M. D. Christian, "Triage," Critical care clinics, vol. 35, no. 4, pp. 575–589, 2019.

[19] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," Manufacturing & Service Operations Management, vol. 5, no. 2, pp. 79–141, 2003.

[20] T. R. Robbins, D. J. Medeiros, and T. P. Harrison, "Does the erlang c model fit in real call centers?" in Proceedings of the 2010 Winter Simulation Conference. IEEE, 2010, pp. 2853–2864.

[21] O. Garnett, A. Mandelbaum, and M. Reiman, "Designing a call center with impatient customers," Manufacturing & Service Operations Management, vol. 4, no. 3, pp. 208–227, 2002.

[22] E. Gijo and N. Balakrishna, "Sarima models for forecasting call volume in emergency services," International Journal of Business Excellence, vol. 10, no. 4, pp. 545–561, 2016.

[23] M. van Buuren, G. J. Kommer, R. D. van der Mei, and S. Bhulai, "EMS call center models with and without function differentiation: A comparison," Operations Research for Health Care, vol. 12, pp. 16–28, 2017.

[24] M. van Buuren, G. J. Kommer, R. D. van der Mei, and S. Bhulai, "A simulation model for emergency medical services call centers," in 2015 winter simulation conference (WSC). IEEE, 2015, pp. 844–855.

[25] M. C. A. van der Burgt, S. Mérelle, A. T. F. Beekman, and R. Gilissen, The impact of COVID-19 on the suicide prevention helpline in the Netherlands. Crisis. 2022.

[26] A. Weatherall, S. Danby, K. Osvaldsson, J. Cromdal, and M. Emmison, "Pranking in children's helpline calls," Australian Journal of Linguistics, vol. 36, no. 2, pp. 224–238, 2016.

[27] R. Whitley, D. S. Fink, J. Santaella-Tenorio, and K. M. Keyes, "Suicide mortality in Canada after the death of Robin Williams, in the context of high-fidelity to suicide reporting guidelines in the Canadian media," The Canadian Journal of Psychiatry, vol. 64, no. 11, pp. 805–812, 2019.

[28] P. Verboon and E. Schulte Nordholt, "Simulation experiments for hot deck imputation," Statistical Data Editing, Methods and Techniques, vol. 2, pp. 22–29, 1997.

[29] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018, pp. 1394–1401.

[30] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting," Data, vol. 4, no. 1, p. 15, 2019.

[31] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R," Journal of statistical software, vol. 27, pp. 1–22, 2008.

[32] L. V. Green, P. J. Kolesar, and W. Whitt, "Coping with time-varying demand when setting staffing requirements for a service system," Production and Operations Management, vol. 16, no. 1, pp. 13–39, 2007.