Recognition and Mapping of Facial Expressions to Avatar by Embedded Photo Reflective Sensors in Head Mounted Display

Katsuhiro Suzuki[*] Keio University
Fumihiko Nakamura[†] Keio University
Jiu Otsuka[‡] Keio University
Katsutoshi Masai[§] Keio University
Yuta Itoh[¶] Keio University
Yuta Sugiura[||] Keio University
Maki Sugimoto[**] Keio University

Abstract

We propose a facial expression mapping technology between virtual avatars and Head-Mounted Display (HMD) users. HMD allow people to enjoy an immersive Virtual Reality (VR) experience. A virtual avatar can be a representative of the user in the virtual environment. However, the synchronization of the the virtual avatar's expressions with those of the HMD user is limited. The major problem of wearing an HMD is that a large portion of the user's face is occluded, making facial recognition difficult in an HMD-based virtual environment. To overcome this problem, we propose a facial expression mapping technology using retro-reflective photoelectric sensors. The sensors attached inside the HMD measures the distance between the sensors and the user's face. The distance values of five basic facial expressions (Neutral, Happy, Angry, Surprised, and Sad) are used for training the neural network to estimate the facial expression of a user. We achieved an overall accuracy of 88% in recognizing the facial expressions. Our system can also reproduce facial expression change in real-time through an existing avatar using regression. Consequently, our system enables estimation and reconstruction of facial expressions that correspond to the user's emotional changes.

Index Terms: H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

1 Introduction

Facial expression is one of the most important features of non-verbal expressions in human communication. In a virtual environment, it is possible to synchronize the facial expression of the virtual avatar with the user's expression using facial capturing technology. However, in case of a head-mounted display (HMD), capturing facial expressions is a challenging due to optical occlusions. In order to synchronize the expression of the avatar with the user's expression in HMD applications, there needs to be a system that can estimate facial expressions while the user is wearing an HMD. Previous researches often extracted facial features from image sequences for recognizing facial expression. This kind of camera-based recognition is not suited for an HMD-based system because the HMD would cover the user's face and restrict his/her body movements.

In this paper, we propose a facial expression recognition technique that works in HMD-based systems through the use of optical sensors. We use photo-reflective sensors, which consist of an infrared light emitter and a phototransistor. Multiple sensors are mounted inside the HMD to measure the distances between the sensors and the skin surface for each facial expression. Two types of neural networks were then trained to create the classifiers: one to determine the closest facial expression through multi-class classification, and another to estimate the degree of similarity through regression. Using these networks, we achieved a smooth transition between the different facial expressions of the avatar.

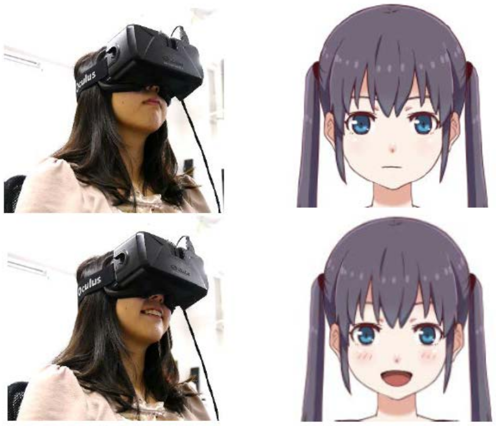

Figure 1: Facial expressions reflected to the avatar (Top: Neutral, Bottom: Happy)

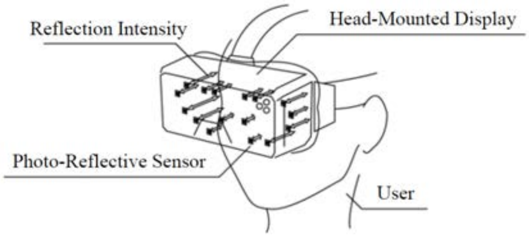

Figure 2: Proposed method

There are many studies that focus on reconstructing realistic facial features and movements with human models. However, if we apply these approaches to unrealistic cartoon characters or nonhuman models that have different facial structures from an actual human, the generated output may look unnatural and hence make it difficult to identify the emotion on its face. Consequently, the facial animation would look choppy. In the above approach, the iden-

[*] e-mail: katsuhirosuzuki@imlab.ics.keio.ac.jp
[†] e-mail: f.nakamura@imlab.ics.keio.ac.jp
[‡] e-mail: jiu@imlab.ics.keio.ac.jp
[§] e-mail: masai@imlab.ics.keio.ac.jp
[¶] e-mail: itoh@imlab.ics.keio.ac.jp
[||] e-mail: sugiura@imlab.ics.keio.ac.jp
[**] e-mail: sugimoto@imlab.ics.keio.ac.jp

2017 IEEE Virtual Reality (VR)
March 18-22, 2017, Los Angeles, CA, USA
978-1-5090-6647-6/17/$31.00 © 2017 IEEE

tified expression is retargeted onto the model in order to realize a smooth transition between expressions. Instead, our system uses a mapping technique that correlates the user's expression with the model's existing animation sequence through machine learning.

The main contributions of our study are as follows: 1) We proposed a facial expression recognition technique that can be used while one is wearing an HMD. We created a system that accommodates the repeated use of HMD by building an inexpensive, lightweight hardware that does not affect the exterior of the HMD. 2) We facilitated the training data collection process by letting the user mimic the avatars a facial expression. 3) We enabled a smooth transition between expressions by training two kinds of neural networks with the user's expressions and the model's animation sequence.

2 Related works

2.1 Camera-Based Approach

The number of studies on facial expression recognition techniques has grown significantly in the field of computer vision. In general, the camera-based techniques for expression recognition are based on detected facial parts. For example, Ying-li et al. developed a system to recognize the subtle movements of facial muscles for six facial expressions by extracting facial features from RGB images [22]. Arumugam et al. classified three emotional states from images by using the Fisher linear discriminant and the singular value decomposition for facial feature extraction, and the radial basic function network for classification [1]. Hommel et al. proposed a method for classifying sequential facial images into seven emotions as well as a method for expressing one's emotions in a one-dimensional emotion space through regression [7]. Both methods are based on active appearance models, which are popular for modeling the human face. In addition to facial expression recognition, there are works that transform 3D face models according to the recognized emotion based on facial features. Itoi et al. extracted features of multiple facial expressions from face images and reproduced the facial expressions by integrating the component ratios of each facial expression [8]. Chin et al. developed a cloning system that transforms emotional intensity-based expression into a low polygon face model [3]. Takahashi et al. developed an application that reflects the user's facial expression to the avatar using RGB images [19]. Thies et al. proposed a method to apply one's facial expression to another in real-time by capturing the facial performances of subjects in the source and target videos [20]. These researches map the expression to a real or virtual face based on facial geometry. The method to capture detailed 3D facial geometry from camera images has also been studied. Pablo et al. reconstructed 3D facial geometry using a monocular camera [5]. Their system associates the 3D face model to the 2D sparse facial features, which selects the keyframe optical flow. Michael et al. presented a system that reconstructs deformable objects including face, arms, and hands using a RGB-D camera [24]. Xavier et al. estimated the expression of a user wearing an HMD from one's mouth. However, this system was not able to recognize many of the expressions since the mouth shape does not change significantly in some expressions. These systems focus on the fidelity in capturing the facial performance [2]. Kyle et al. proposed a method by which an HMD wearer can control the facial expression of the avatar using a convolutional neural network trained on mouth-region image and animation data [13]. Justus et al. performed real-time gaze-aware facial reenactment using an RGB-D sensor with a frontal view and an infrared camera located inside the head-mounted display [21]. Since these methods target high-resolution avatars, high computational costs and expensive hardware are required.

2.2 Contact-based Approach

Techniques for recognizing facial expression are explored in other fields as well. Jocelyn et al. proposed a system to identify specific facial expressions by sensing facial movement [16]. It is one of the first works to apply the recognition technique to a wearable device. Partala et al. estimated two emotional states (smile and frown) from electromyographic (EMG) activity of two facial muscles [14]. Gruebler et al. developed a wearable device composed of two electrode pairs in order to detect facial EMG signals for recognizing positive expressions [6]. Li et al. developed a technique that combines the camera-based approach and contact-based approach[9]. They use an RGB-D camera and strain gauges that make contact with the face. The data from the strain gauges and camera are combined and applied to the 3D face model. However, the strain gauge requires a stable contact to the face. In order to solve this problem, we propose a robust facial expression recognition system by capturing facial geometry with non-contact optical sensors.

2.3 Interaction Using Optical Sensors

Optical sensors can detect the presence of an object and its position by means of radiating infrared light and receiving the reflected light. Sugiura et al. measured the deformation ratio of the cloth by using optical sensors based on the fact that the light transmittance changes according to the mesh size of the cloth [18]. Also, there are studies that identifiy the status of the body by utilizing the distance values between the optical sensor and the human skin surface. Ogata et al. allow for the identification of various interactions by mounting an optical sensor onto a bracelet-type device [12]. Fukumoto et al. developed a smile/laughter-based life logging system using two optical sensors [4]. Masai et al. estimated facial expressions with an eyeglass-type device where a number of photosensors are placed on the inner side of the eyeglass. This approach is also effective for HMDs because large objects cannot be attached inside an HMD. Thus, we have adopted the sensing technique that uses optical sensors for our purpose [10] .

3 Facial expression mapping by embedded optical sensors on a head-mounted display

Multiple optical sensors can be attached to the interior of an HMD along with a flexible circuit board due to the fact that the sensors are small enough and have a small power consumption. The sensor measures the distance between the sensor and the surface of the skin, which slightly fluctuates as the facial expression changes. The data obtained from the sensors allows the neural network to estimate the current facial expression. Likewise, the system can estimate continuous expression change using regression. In addition, the proposed method preprocesses the sensor values and automatically collects the training datasets (a set of sensor values and labels representing the facial expression) in order to learn the neural network. Also, it can reflect the user's facial expressions onto those of the avatar. The process of preprocessing the sensor values for robust expression recognition is described in Section 3.1, the automatic collection of the training datasets through the imitation of the virtual avatar is described in Section 3.2, the network architecture of multi-class classification is described in Section 3.3, the network architecture of the regression neural network is described in Section 3.4, the process of correlating the user's facial expression with the avatar's expression using two classifiers is described in Section 3.5.

3.1 Preprocessing

If raw sensor values are used for neural network training, the accuracy rate of expression recognition would significantly fall when the position of the HMD shifts even a little. In order to accommodate the shift of the HMD's position, we took the difference between the sensor values of the neutral expression and every other expression, which in fact is fairly constant. Hence, we record the sensor values of the neutral expression as the baseline. Then, we ask the subject to move the face muscles randomly to record the maximum and minimum values for each sensor. The sensor value has a non-linear relationship to the distance between the sensor and the surface. To achieve a linear relationship between these variables, the sensor values are compensated as shown in Fig.3.

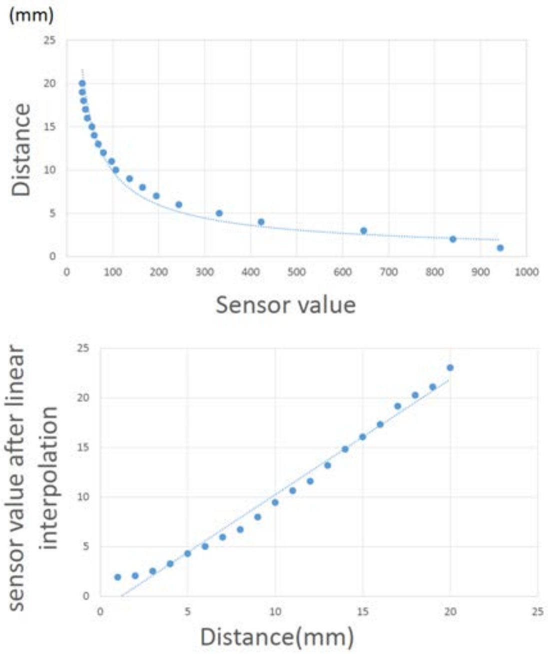

Figure 3: Linear interpolation. (top) Relationship to the sensor value and the distance between a sensor and the face surface. (bottom) Linear interpolation using the approximate formula.

3.2 Training Data Collection by Mimicking Virtual Characters

We automated the training phase in order to reduce the time and effort of the user. We created a virtual avatar that makes five expressions (1: Neutral, 2: Happy, 3: Angry, 4: Surprised, 5: Sad) at constant intervals. During the training phase, the user is instructed to mimic the facial expression of the avatar. In the meantime, our system automatically collects the sensor data and supervised data that corresponds to the facial expression label or intensity. In order to avoid collecting sensor data with undesired supervised data, the avatar changes its facial expressions with enough transition time. As we use two neural networks with multi-class classification and regression, our system collects a different dataset for each neural network. Therefore, the avatar changes its facial expressions in two ways. For the multi-class classification dataset, our system collects the sensor data and facial expression label while the avatar changes its expressions to the five different expressions in order. For the regression dataset, our system collects the intensity data (0 – 1) in addition to the sensor data and facial expression label, while the avatar changes its facial expression from neutral to the other four expressions with smooth transition. For example, our system collects the data while the avatar's expression changes from neutral to happy, and from neutral to angry.

3.3 Multi-Class Classification of Facial Expression

We classify the five facial expressions with a neural network. The output is determined by propagating the input value sequentially through the node. Since the neural network of the proposed method has a fully connected layer, the value of the input node z is determined by summing the product of the previous layer node x and weight w. The output of the node is determined by applying the activation function f to the input of the node.

$$z_j = f(\sum w_{ji} x_i + b_j)$$

In the network for multi-class classification, the activation function of the input layer and hidden layer uses the rectified linear unit (ReLU)[11].

$$f(u) = max(u, 0)$$

ReLU works better than the hyperbolic tangent function or softplus function [23]. We use backpropagation for neural network adjustment [15]. For the error function E, we used the cross-entropy function, which is the most common method with low computational cost in likelihood estimation.

$$E(w) = \sum_{n=1}^{N} \sum_{k=1}^{K} d_{nk} \log y_k(x_n; w)$$

The activation function of the output layer is the softmax function. Each dimension of the output layer corresponds to each of the five basic facial expressions. We choose the facial expression that has the highest value of the corresponding dimension of the output layer, i.e. the highest likelihood.

$$f(u) = \frac{\exp(u_k)}{\sum \exp(u_j)}$$

The neural network needs sets of training data and teaching signals for training. In the proposed method, we selected five basic facial expressions (neutral, happy, angry, surprised, sad) as the recognition target. We adopted a three-layer feedforward neural network with 16 inputs, 100 hidden, and 5 outputs. The number of inputs represents the number of sensors, and the number of outputs represents the number of facial expressions. The number of hidden was chosen to show the best accuracy through preliminary experiments. The learning rate was set to 0.001 and epoch was set to 100. With these parameter settings, the neural network converges with all subjects. The final loss value was 0.01. We used dropout, a function that learns without some nodes of the hidden layer randomly in order to prevent over-fitting [17]. In the case of training 40,000 values of data (16 sensors x 5 facial expressions x 100 frames x 10 trials), the neural network training takes about 30 seconds when using a computer with the following specifications: Intel Core i7-4770 processor, NVIDIA GTX 760, and 8GB memory. The output of the neural network was used to select the regression neural network.

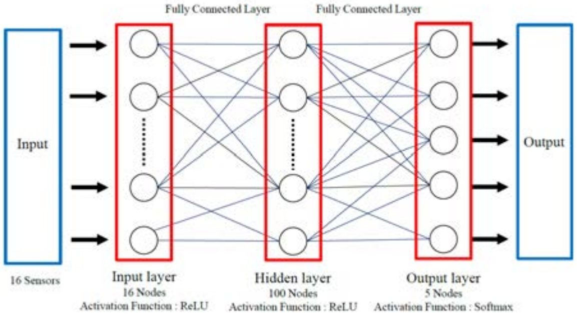

Figure 4: Network architecture for multi-class classification

3.4 Estimation of Intermediate Expression by Regression

The classifier of multi-class classification is not guaranteed to change the output linearly as the subject's facial expression changes linearly. Therefore, we made a classifier with a regression neural network to identify the continuous change in facial expression. The architecture of the neural network is the same as multi-class classification except for the activation function of the output layer, the error function, and the number of output nodes. The output of the regression neutral network is similarity of facial expression between neural and other basic facial expressions, so only one output node is required. For the error function E, we used the mean squared error function, which is often used in cases where the neural network outputs a continuous value.

$$E(w) = -\frac{1}{2}\sum ||d_n - y_n||^2$$

The activation function of the output layer is a hyperbolic tangent function. The number of classifiers is the same as the number of facial expressions except for neutral, and we selectively used it by a result of multi-class classification.

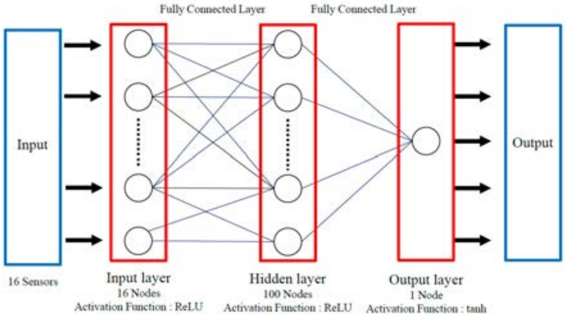

Figure 5: Network architecture for regression

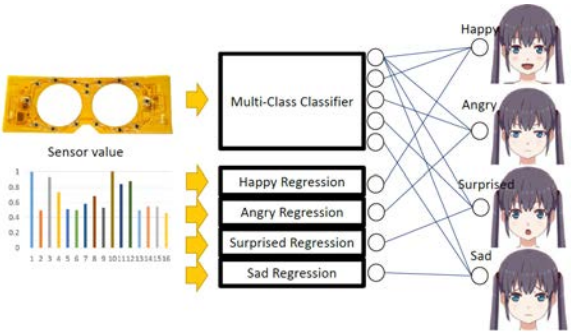

Figure 6: Overview of the system. The outputs are the most similar facial expressions and the similarity between the neutral and most similar expression.

3.5 Transforming Facial Expression into Virtual Avatars

The proposed method can estimate the facial expression of the HMD user by using multi-class classification and regression of the neural network. Therefore, we have developed an application that reflects the estimated facial expression on the avatar's face . First, the subject wears the HMD and sets the baseline, maximum value, and minimum value. Then, we record the datasets that are necessary for multi-class classification and regression by mimicking the avatar. Thereafter, it is possible to reflect the user's facial expression to the avatar using the classifier created by the neural network. The collection of datasets and creation of the neural network take a few minutes. It is possible to recognize facial expressions by preprocessing the sensor value even if the HMD is removed in the middle of the experiment. When the user wants to wear the HMD again, it is possible to operate without any problem if the baseline is reset. Also, the datasets can be reused for the same individual. When the wearer's facial expression is reflected to the avatar, the avatar's facial expression changes smoothly because of sensor noise. We were able to change the facial expression smoothly by averaging the output of the regression.

4 Implementation

4.1 Hardware

We designed a flexible printed circuit board (PCB) that can easily be mounted in the interior of the Oculus Rift Development Kit 2 (DK2). We installed 16 photo reflective sensors (Kodenshi SG-105) on the flexible PCB. We placed 14 optical sensors around the eyes and two optical sensors on the bottom part of the HMD device to measure the movement of the user's cheeks. As for the reasons for choosing this layout, we found that the movement of skin at these positions are prominent and that the sensor values scatter when facial expressions change to an extent that discrimination is possible. Individual differences would make it impossible to place the sensors over muscles. In fact, we ensure usability by not taking the position of muscles into account. Each sensor is connected to Microchip 16F619. The sensor values are passed to a computer by a serial communication via USB.

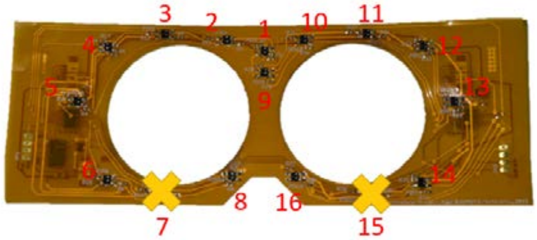

Figure 7: Flexible circuit board. The 7th and 15th sensors were reconnected at the bottom of the HMD

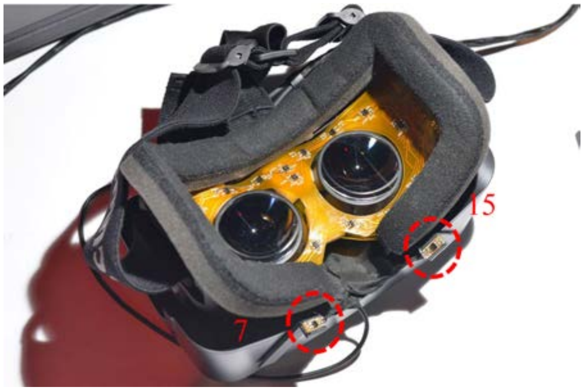

Figure 8: Interior of the HMD.The circuit board is connected via USB port

4.2 Software

We implemented facial expression mapping between the user and the virtual avatar using Unity, iClone, and Live2D. Unity is a game engine that can draw 3D models easily. iClone and Live2D provide virtual avatars that can animate on the engine. We are able to manipulate the facial parameters of the avatar in this environment. A list of parameters that can be modified is shown in Table 1. Also, the gyro sensor mounted on the Oculus Rift allows us to obtain the rotation angles of the head. The neural network was implemented with Python and a machine learning library called Chainer. Also, we made a user interface to visualize the sensor value and the state of our system using Processing. Processing gets the sensor values via serial communication. Data such as sensor values, identification results, and keyboard inputs were transmitted between Processing, Unity, and Python via User Datagram Protocol (UDP) communication.

Table 1: Parameters of the Live2D avatar

Name	Range of Value	N	H	A	Sur	Sad
EYE_OPEN	0 ˜ 2	1	1	0.8	2	0.8
EYE_FORM	-1 ˜ 1	0	0	0	0	-1
EYE_BALL_FORM	-1 ˜ 1	0	0	0	-1	0
MOUTH_FORM	-1 ˜ 1	0	1	-1	-1	-1
MOUTH_OPEN_Y	0 ˜ 1	0	0.5	0	1	0
BROW_ANGLE	-1 ˜ 1	0	0	-1	0	-1
BROW_FORM	-1 ˜ 1	0	0	-1	0	-1
BROW_Y	-1 ˜ 1	0	0	0	1	-1
TERE (Awkward)	-1 ˜ 1	0	1	0	0	0

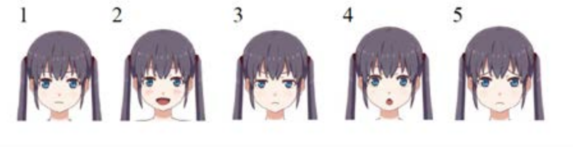

Figure 9: Basic facial expressions of a Live2D avatar (1. Neutral, 2. Happy, 3. Angry, 4. Surprised, 5. Sad).

5 Experiment

5.1 Experiment 1: Precision experiment of facial expression recognition

Participants were instructed to imitate and maintain the facial expression specified by the avatar (neutral, happy, angry, surprised, sad) shown in the HMD. Before data collection, we first let the participants practice in order to familiarize themselves with moving their face. The expression of the avatar changed every 100 frames (5 facial expressions * 100 frames * 10 trials = 5000 sets of data for each subject). Once the participants were ready, 10 trials were conducted. They would take the HMD off after every trial. We assigned the first five trials of data for training the network, and the rest for verifying the accuracy of facial expression recognition. The participants included 9 men and 1 woman in their 20s. The training and identification processes were done for each individual. The recognition accuracy values of the facial expressions are shown in Table 2.

Table 2: Recognition accuracy of facial expressions

Actual \Predict	Neutral	Happy	Angry	Surprised	Sad
Neutral	95.74%	1.12%	0%	1.96%	1.18%
Happy	0.74%	98.86%	0.04%	0.36%	0%
Angry	3.94%	0.08%	79.98%	0%	16%
Surprised	0%	0%	0.44%	92.12%	7.44%
Sad	9.36%	02%	12.08%	2.16%	76.2%

As a result, the average accuracy of facial expression recognition was 88%. The recognition accuracy of Happy was high for all subjects compared to other expressions due to the distinctive movement in the cheeks when one smiles. The recognition accuracy of Angry and Sad were low and easily mistaken with one another in many of the cases. The accuracy rate of each individual is shown in Fig.10.

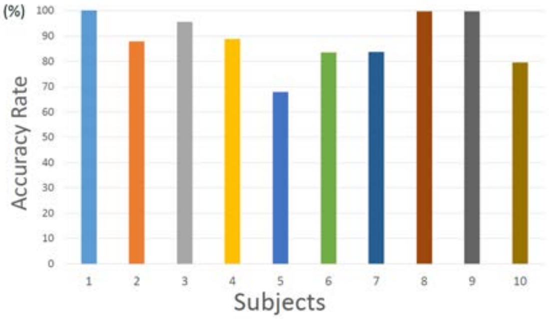

Figure 10: Recognition accuracy of individuals of Experiment 1

5.2 Experiment 2: Precision Experiment of the Automatic Data Collection by the Avatar

We first verified whether the sensor values change when the subject mimics each facial expression of the avatar. We performed a principal component analysis on all sensor values. The first and second principal components are shown in Fig.11. We confirmed that the sensor values undoubtedly change when the subject mimics each expression of the avatar. Also, we confirmed that the subjects were given sufficient time to change their expression.

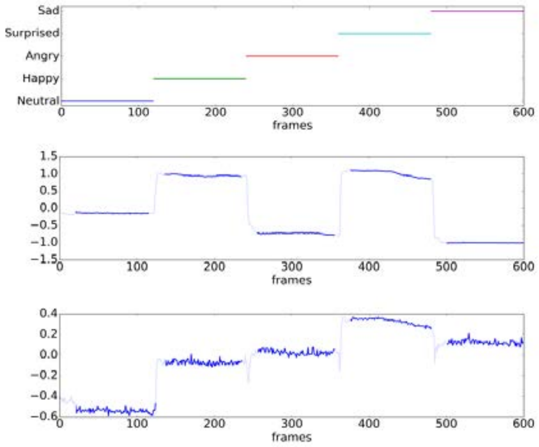

Figure 11: Transition of the sensor value when the user mimics the facial expression of the avatar. (top) Facial expression of the avatar. (middle) First principal component. (bottom) Second principal component.

Second, we verified whether the automatic data capture system can collect the training data only by letting the subject imitate the avatar. The procedure used was the same as Experiment 1. Learning and identification were done for each subject which consisted of 10 men in their 20s. The recognition accuracy values of the facial expressions are shown in Table 3.

The recognition accuracy was 84% on average. Although the accuracy decreased by 4%, the accuracy was almost the same as when the training data was collected by instruction. The recognition rates of each individual are shown in Fig.12.

Table 3: Recognition accuracy of facial expressions using automatic data capture.

Actual \ Predict	Neutral	Happy	Angry	Surprised	Sad
Neutral	96.36%	0.22%	0.16%	3.26%	0%
Happy	0.44%	92.98%	3.1%	3.48%	0%
Angry	9.96%	2.98%	78.5%	0.28%	8.28%
Surprised	7.08%	5.16%	1.98%	85.5%	0.3%
Sad	10.02%	0.96%	19.56%	3.26%	66.2%

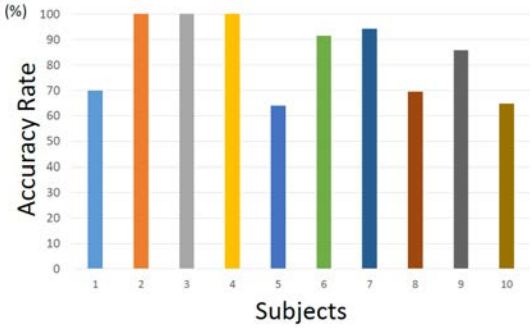

Figure 12: Recognition accuracy of individuals of Experiment 2.

5.3 Experiment 3: Verification of Intermediate Facial Expression Estimation by Regression

We first verified whether the sensor values change continuously when the subject mimics each continuous facial expression. Then, we performed a principal component analysis on all sensor values. The first and second principal components are shown in Fig.13. Although the gradient is not constant, we confirmed that the sensor values change continuously.

Next, we verified whether the regression neural network can estimate continuous facial expression changes by using an application that displays an avatar changing its expression gradually. We obtained six sets of data (5 facial expressions * 100 frames * 6 sets) to create a multi-class classifier. The first five sets of data were used to create the regression neural network, and the last set of data was used for testing. The results are summarized in Fig.14. The second graph shows the result for multi-class classification. The rest of the graphs show the results for regression merged with the multi-class classifier.

5.4 Experiment 4: Comparison of facial expressions between user and avatar

In order to compare the facial expressions of the user and the avatar, we made an experimental device that can simultaneously provide the sensor values and the facial image of the user by removing the display of the Oculus Rift Development Kit 1 (DK1). The sensors were attached in the same position as the proposed device. The experimental device is shown in Fig.16.

We reflected the facial expression to the avatar using this device to make a classifier in the same procedure as Experiment 3. We prepared male and female avatars, and observed the four facial expressions (happy, angry, surprised, and sad). The result is shown in Fig.17. Since our goal was not to synchronize the facial geometry, some facial parts of the avatar were not identical to the user's expression. However, we were able to read the subject's emotion from the facial expression of the avatar to a certain extent. This shows that our method provides appropriate mapping between the user's and avatar's expression.

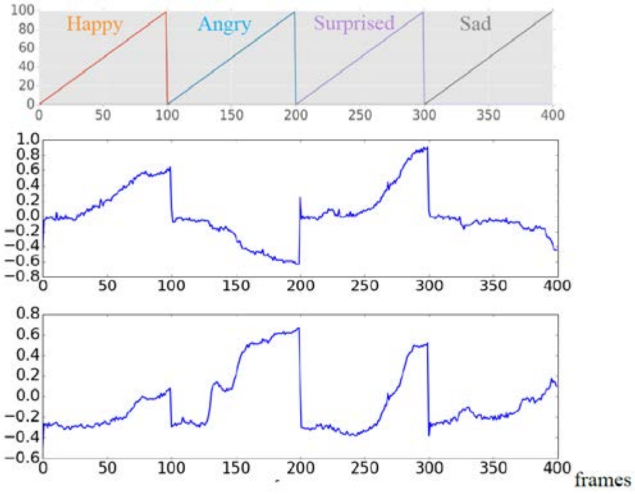

Figure 13: Continuous transition of facial expression. (top) Facial expression change of the avatar. (middle) First principal component. (bottom) Second principal component.

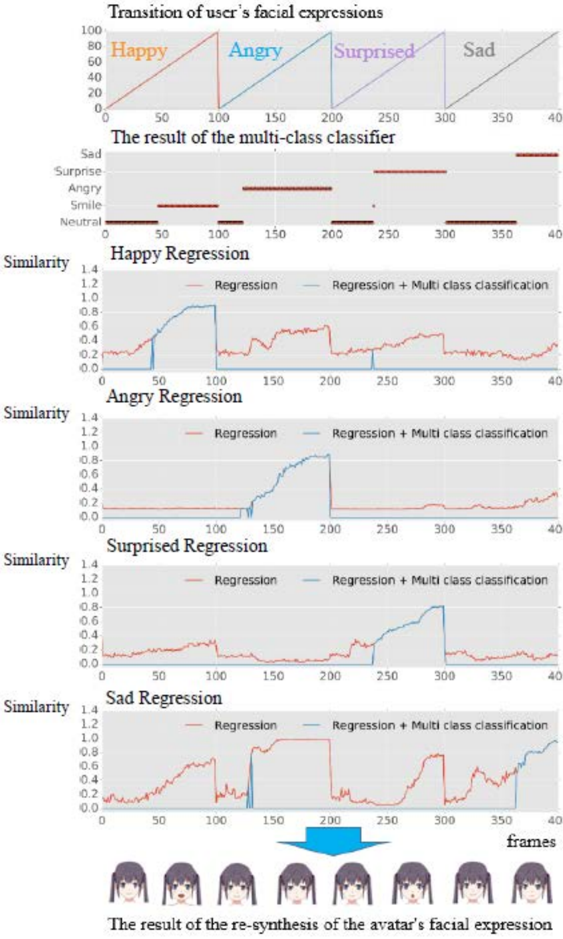

Figure 14: The changes in output due to the continuous change of facial expressions.

6 Discussion

In Experiment 1, the overall recognition accuracy was 88%. Considering the fact that the participants intermittently took off the

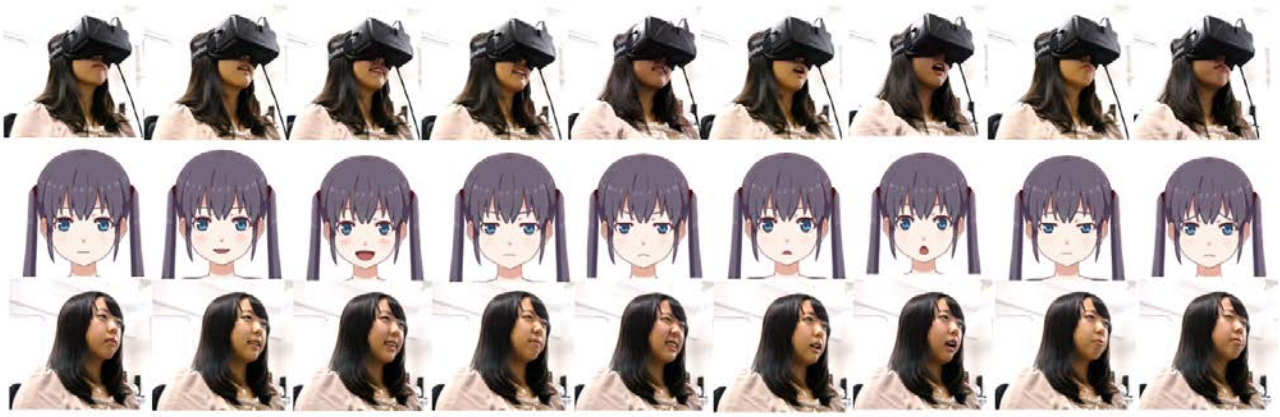

Figure 15: Continuous transition of facial expressions. (top) User's facial expression covered by the HMD. (middle) Facial expressions of the avatar. (bottom) Reproduction of the facial expressions.

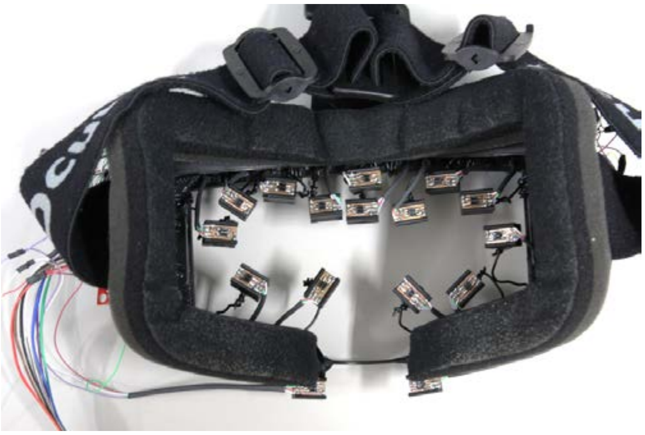

Figure 16: Uncovered HMD for Experiment 4.

HMD during the experiment, the accuracy seems to be good in practical conditions. Fig.18 shows an example of the sensor value distributions of Happy in different trials. The recognition accuracy of Happy showed a good performance for most of the participants. In particular, the movement of the cheeks was prominent compared to other expressions. The miscategorization that happened between Anger and Sadness could be due to the difficulty in reproducing these expressions in the same way. It may also be because some of the participants tended to knit their eyebrows for both emotions, and because both the emotions represent negative feelings.

Experiments 1 and 2 confirmed that the recognition rate of each subject has a large variation. Subjects with low recognition accuracy had a characteristic that some sensor values did not change even if they changed the expression. The reason is that some of the sensor values go out of the detection distance range from the skin surface (1– 20 mm), even if the position of the HMD was adjusted using bands. Throughout Experiments 1 and 2, the data collection method of mimicking was almost as effective as the manual instruction. It is assumed that the small difference of accuracy rate comes from the relative difficulty of mimicking the facial expressions of the avatar compared to creating the facial expression of the instructed emotion in the participant's mind. The result of regression in Experiment 3 shows a sudden increase in several points since it is difficult to in practice change expressions perfectly in a linear fashion. In addition, since subjects tended to change their expressions rapidly soon after realizing that the expression of the avatar had changed, it is inevitable that the result of the regression deviates from a linear relationship. The small noises in the graph may result from tremors in the subject's facial muscles as they were exhausted from changing their expressions slowly. The proposed method processes low-pass filters or the moving average method in the software in order to reduce the sensor noise. Based on our experiments, we confirmed that the proposed method is able to recognize facial expressions and to reproduce continuous facial expression change. The facial expressions of the avatar did not perfectly correspond to the subject's expressions in Experiment 4. However, the proposed method can express user's complicated facial movement by mapping the facial expression to an existing animation model, though it is difficult for some people to mimic exaggerated 2D/3D avatar's expressions. If the facial expression is reflected to

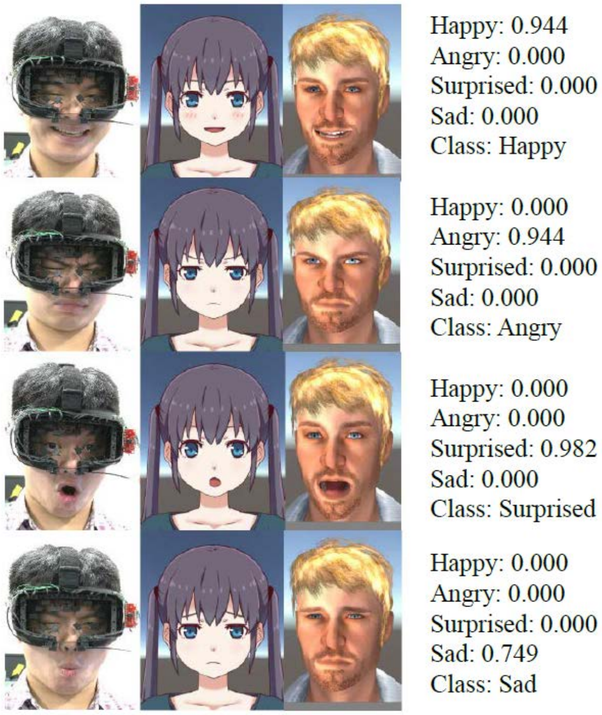

Figure 17: The result of Experiment 4 (top to bottom: Happy, Angry, Surprised, Sad).

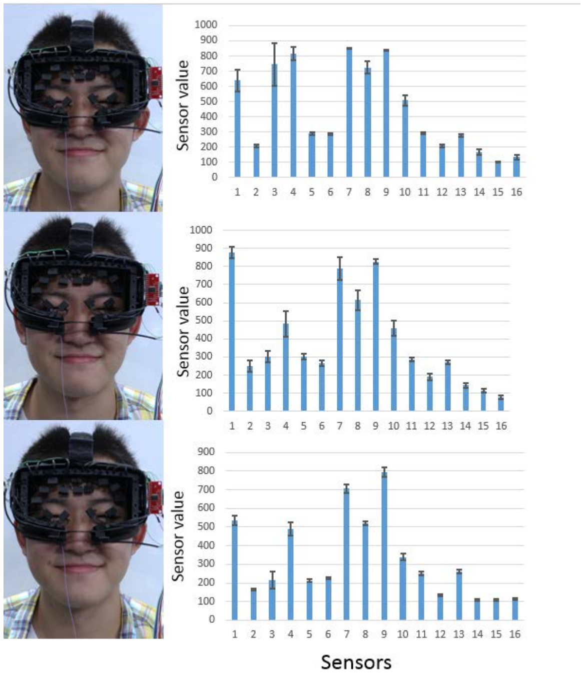

Figure 18: Sensor value distributions upon taking off and putting on the HMD.

the avatar, the impression of the avatar on users is also important. Reading the user's emotion from the avatar's facial expression is expected to enhance the usability of the proposed method.

7 Limitation and future work

Sensor values differ among individuals even when they make the same facial expression due to differences in their facial features. Therefore, calibration is necessary for each individual. In the future, we wish to collect data from various people of different races and to expand the system so that it can learn the variations of sensor values that are common for all people. Although we chose to use a neural network in the proposed method with reference to [10], the training speed may improve by using a support vector machine or decision tree, considering the fact that the sensor data we used has simple features. Hence, we would like to compare various machine recognition methods with respect to recognition accuracy and training speed. Furthermore, how natural the animation looks is dependent on the original animation. The model used in this study offers a function that can adjust the parameters of the model in detail. However, the animation is still not natural enough because it has been created by someone who is not proficient in animation. When the wearer makes a facial expression that the neural network has not learned, it cannot resynthesize the estimated look from the learned facial expressions. In addition, we have not yet reproduced the wearer's mouth movements when speaking since prototype implementation only uses two sensors to capture the movements of the mouth. In the future, we would like to increase the number of sensors around the mouth so that we can incorporate facial animation for the mouth region into our system. Further investigation is also needed with regard to finding a better arrangement of the sensors as well as which type of sensor to use. Through facial expressions, people express various kinds of emotions in real life. In the proposed method, we focused on five basic facial expressions, but we look forward to investigating whether our system can identify more kinds of facial expressions. Moreover, we would like to investigate whether we can recognize facial expression changes that do not express emotion, such as winking.

8 Conclusion

We proposed a facial expression mapping technique between the user and the avatar using embedded optical sensors and machine learning. By making use of the characteristics of optical sensors, we can build the system without adding anything onto the exterior of the HMD. Since the sensors do not require physical contact with the user's face, we made the system to work even when the HMD is shifted or reattached through the use of calibration based on neutral expression. The dataset collected by mimicking an avatar shows nearly the same recognition rate as the dataset collected with manual instruction. Also, there were big variations among individuals because the sensor values of some individuals went beyond the detection distance of the sensors. The arrangement and choice of sensor requires further investigation. By combining classification and regression neural networks, we reproduced smooth transitions in the animation of facial expression change. As a result of our integration, we were able to create natural facial expressions of an avatar that correspond to the user's facial expressions.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16H05870.

References

[1] D. Arumugam and S. Purushothaman. Emotion classification using facial expression. *International Journal of Advanced Computer Science and Applications*, 2(7), 2011.

[2] X. P. Burgos-Artizzu, J. Fleureau, O. Dumas, T. Tapie, F. LeClerc, and N. Mollet. Real-time expression-sensitive hmd face reconstruction. In *SIGGRAPH Asia 2015 Technical Briefs*, page 9. ACM, 2015.

[3] S. Chin and K.-Y. Kim. Emotional intensity-based facial expression cloning for low polygonal applications. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(3):315–330, 2009.

[4] K. Fukumoto, T. Terada, and M. Tsukamoto. A smile/laughter recognition mechanism for smile-based life logging. In *Proceedings of the 4th Augmented Human International Conference*, AH '13, pages 213–220, New York, NY, USA, 2013. ACM.

[5] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 158:1–158:10, November 2013.

[6] A. Gruebler and K. Suzuki. Design of a wearable device for reading positive expressions from facial emg signals. *Affective Computing, IEEE Transactions on*, 5(3):227–237, 2014.

[7] S. Hommel and U. Handmann. Aam based continuous facial expression recognition for face image sequences. In *Computational Intelligence and Informatics (CINTI), 2011 IEEE 12th International Symposium on*, pages 189–194. IEEE, 2011.

[8] I. Kiyoaki, M. Yasushi, and K. Yukio. Intelligent coding of facial image using neural network and morphing. *IEEJ Transactions on Electronics, Information and Systems*, 120(8-9):1165–1171, 2000.

[9] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), July 2015.

[10] K. Masai, Y. Sugiura, M. Ogata, K. Kunze, M. Inami, and M. Sugimoto. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, pages 317–326, New York, NY, USA, 2016. ACM.

[11] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010.

[12] M. Ogata, Y. Sugiura, Y. Makino, M. Inami, and M. Imai. Senskin: Adapting skin as a soft interface. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 539–544, New York, NY, USA, 2013. ACM.

[13] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Trans. Graph.*, 35(6):221:1–221:14, Nov. 2016.

[14] T. Partala, V. Surakka, and T. Vanhala. Real-time estimation of emotional experiences from facial expressions. *Interacting with Computers*, 18(2):208–226, 2006.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Leraning representations by back-propagation errors. *Nature*, pages 533–536, 1986.

[16] J. Scheirer, R. Fernandez, and R. W. Picard. Expression glasses: A wearable device for facial expression recognition. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, pages 262–263, New York, NY, USA, 1999. ACM.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[18] Y. Sugiura, G. Kakehi, A. Withana, C. Lee, D. Sakamoto, M. Sugimoto, M. Inami, and T. Igarashi. Detecting shape deformation of soft objects using directional photoreflectivity measurement. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 509–516, New York, NY, USA, 2011. ACM.

[19] K. Takahashi and Y. Mitsukura. Eye blink detection using monocular system and its applications. In *RO-MAN, 2012 IEEE*, pages 743–747. IEEE, 2012.

[20] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6):183, 2015.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *CoRR*, abs/1610.03151, 2016.

[22] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):97–115, Feb. 2001.

[23] Y. B. Xavier Glorot, Antoine Bordes. Deep sparse rectifier neural network. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2015.

[24] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4):156, 2014.