International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2019)
Y. Kakehi and A. Hiyama (Editors)

Automatic Labeling of Training Data by Vowel Recognition for Mouth Shape Recognition with Optical Sensors Embedded in Head-Mounted Display

Fumihiko Nakamura[1] , Katsuhiro Suzuki[1] , Katsutoshi Masai[1] , Yuta Itoh[2,3] , Yuta Sugiura[1] and Maki Sugimoto[1]

[1]Keio Univesity, Japan
[2]Tokyo Institute of Technology, Japan
[3]RIKEN Center for Advanced Intelligence Project, Japan

Abstract

Facial expressions enrich communication via avatars. However, in common immersive virtual reality (VR) systems, facial occlusions by head-mounted displays (HMD) lead to difficulties in capturing users' faces. In particular, the mouth plays an important role in facial expressions because it is essential for rich interaction. In this paper, we propose a technique that classifies mouth shapes into six classes using optical sensors embedded in HMD and gives labels automatically to the training dataset by vowel recognition. We experiment with five subjects to compare the recognition rates of machine learning under manual and automated labeling conditions. Results show that our method achieves average classification accuracy of 99.9% and 96.3% under manual and automated labeling conditions, respectively. These findings indicate that automated labeling is competitive relative to manual labeling, although the former's classification accuracy is slightly higher than that of the latter. Furthermore, we develop an application that reflects the mouth shape on avatars. This application blends six mouth shapes and then applies the blended mouth shapes to avatars.

CCS Concepts

• ***Human-centered computing*** → ***Human computer interaction (HCI);***

1. Introduction

VR systems allow their users to communicate via avatars in various scenarios. In such a case, the facial expression is essential because it conveys emotions and intentions, among others. Computer vision techniques can be used to capture facial expression in many cases; however, facial expressions under HMD cannot be easily captured due to facial occlusions caused by the display. Several solutions have been developed to overcome this issue. One remarkable method is using embedded optical sensors around the eye region and adopting machine learning techniques to recognize facial expressions [KFJ*17]. However, given that the sensors are allocated around the eye region, the system recognizes only limited movements of the mouth. The mouth is important for understanding expressions [YMM07], especially for Westerners.

In this paper, we introduce a system that recognizes the mouth shape of the HMD user with optical sensors. We adopt two types of optical sensors: a photo-reflective sensor, and a position sensitive detector (PSD). These sensors measure the distances between them and the skin surfaces surrounding the mouth. We collect optical sensor values for each mouth shape and then label the sensor values using vowels, which are detected from speech. We train classifiers using this labeled sensor values. We build an HMD-based prototype that measures mouth shape and develop an application that reflects the user's mouth shape on an avatar. This application blends mouth shapes according to belonging probabilities for the classes. Figure 1 shows this application applying a blended mouth shape to an avatar.

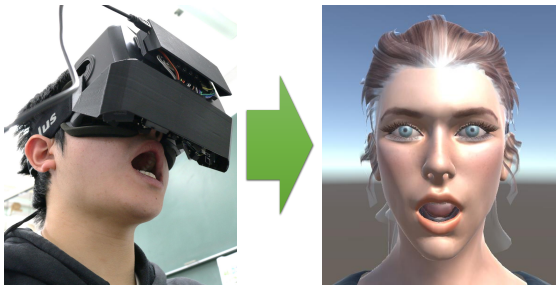

Figure 1: *Reflecting Mouth Shape of HMD User to Avatar*

Many studies on capturing facial performance use cameras or optical sensors. Camera-based techniques detect facial gestures ac-

DOI: 10.2312/egve.20191274

curately, but they have difficulty integrating to HMD-based systems because of limitations such as weight, hardware cost, and high computational cost. Meanwhile, optical sensors are lightweight, low-cost, and capable of recognizing gestures with low-dimensional data. Therefore, they are suitable for wearable devices, such as HMD. In particular, integrating machine learning with sensing with optical sensors enables powerful gesture recognition [MSO*16]. Many of such recognition requires the manual collection of training data, which is time-consuming. To address this problem, we acquire labels by speech recognition to gather training data automatically.

The main contribution of our paper is as follows:

- We developed a technique that recognizes mouth shapes while the user is wearing an HMD. We built a mouth shape sensing device that is lightweight, low-cost, and unaffected by the facial occlusion caused by HMD.
- We collected training data automatically by speech recognition. We obtained labels for training data by integrating vowel recognition with a mouth shape measurement technique.
- We built an application that can reflect the user's various mouth shapes on avatars by synthesizing bases of the mouth shape. The parameters of bases are blended according to belonging probabilities to reproduce multiple mouth shapes.

2. Related Work

We review previous works on capturing facial performance and highlight wearable systems with embedded sensors. This section describes an overview of sensing approaches, such as audio, camera, contact sensors, and optical sensors.

2.1. Audio-based Approach

Speech has been used to produce an animation of lip movements. Speech consists of phonemes, the smallest unit of speech that makes sense as a language. A mouth shape and a tongue position determine a phoneme. A mouth shape corresponded to a lip position. Therefore, phoneme detection leads to estimation of lip movements [GP05]. Oculus Lipsync [Uni] recognizes the speech sounds of HMD wearers and matches the lip movements of avatars with the speech in real time. The audio-based approach is based on the speech, and thus does not work without any voice.

2.2. Camera-based Approach

One of the most popular approaches to capturing facial movement is the use of cameras. Focusing on HMD users' facial movement recognition, embedded camera approaches have been explored. For example, Hickson et al. [HDS*19] classified facial expressions from images of an eye camera embedded in HMD. Olszewski et al. [OLSL16] developed a system that reconstructs the facial geometry of the user using an HMD with both eye cameras and a mouth camera. However, this approach requires high computational power and expensive hardware because it targets high-fidelity avatars.

2.3. Contact-based Approach

Contact sensors can detect facial movements through muscles and skin deformation. For example, Gruebler et al. presented a wearable device to recognize positive facial expressions from electromyography [GS14]. Li et al. proposed a system that reconstructs facial geometry from both strain gauges and an RGB-D camera attached to an HMD [LTO*15].

Contact sensors are suitable for wearable devices because of their compactness and fulfillment of the required contact with surfaces. However, contact-based sensing techniques depend on the condition of the contact with a surface, and there are concerns about comfort while the device is mounted.

2.4. Measurement using Optical Sensors

Optical sensors have been deployed to wearable devices because of their capability to sense gestures. Some interfaces using optical sensors focus on the HMD wearer. Sakashita et al. developed a mask-type interface that transmits human action to puppetry [SMK*17]. Their system used optical sensors to detect the lower lip position and classifies three mouth states (closed, partly open, open). Suzuki et al. built an HMD-based system that recognizes five facial expressions of the HMD wearer [KFJ*17]. Their system used machine learning to recognize various facial expressions. In their system, optical sensors detected the deformation around the eyes. However, this system has difficulties in measuring the mouth shape because it focuses on the eye region.

Combining machine learning to measuring by optical sensors enables the detection of various gestures [MSO*16] but requires a training process. To automate this process, Suzuki et al. requested individuals to imitate the facial expressions of avatars and collected training data [KFJ*17]. However, their system can label training data incorrectly because users can make facial expressions that differ from those of avatars.

As mentioned above, previous studies have two limitations, namely, recognition methods of mouth shapes and labeling methods of training data. We measure and recognize the mouth shape with optical sensors. We adopt machine learning to recognize various mouth shapes. In training, we recognize vowels to give correct labels to training data.

3. Mouth Shape Recognition by Embedded Optical Sensors in HMD

This section provides an overview of our mouth shape recognition system. Our system consists of three techniques, namely, mouth shape classification, data labeling using vowel recognition, and mouth shape reproduction. Optical sensors detect the distances between them and skin surfaces. Optical sensor values are labeled with vowels recognized from speech and are learned to classify the mouth shapes. In reproducing the mouth shapes, we blend multiple mouth shapes according to the belonging probabilities. This reproduction approach is similar to that in a previous research [IYY01]. Figure 2 shows the flow of mouth shape recognition by optical sensors and labeling of training data by vowel recognition. Section 3.1

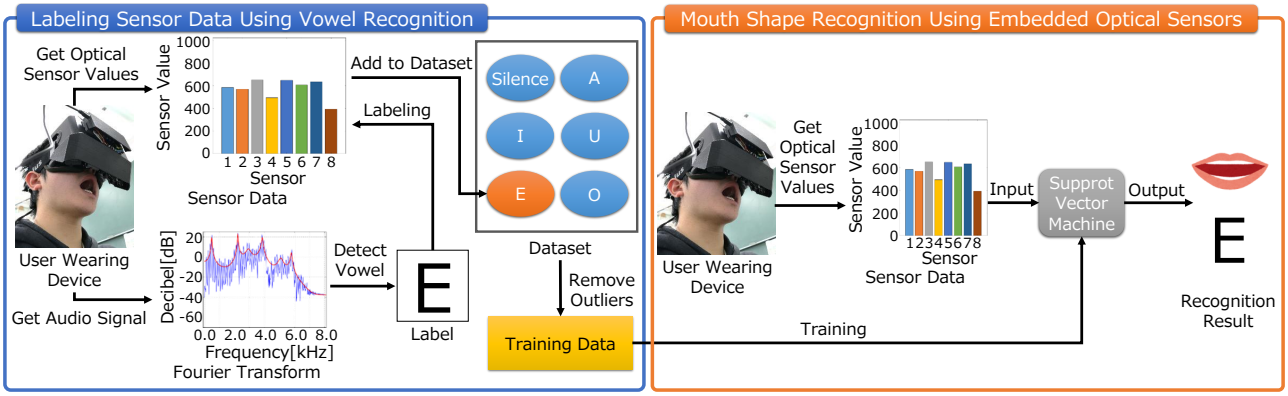

Figure 2: *Mouth Shape Recognition using Optical Sensors and Labeling Method of Training Data using Vowel Recognition*

describes the mouth shape measurement and the classification process. Section 3.2 describes the labeling technique of the training data using vowel recognition. Section 3.3 describes the blending method for the mouth shapes.

3.1. Mouth Shape Classification by Embedded Optical Sensors

Our system measures mouth shapes by an approach similar to that in [MSO*16]. We use embedded optical sensors to achieve an optimized sensor allocation and low computational cost. The skin deforms as the mouth muscles move. Optical sensors capture this deformation by detecting the distance between them and the skin surfaces. These distances are different for each mouth shape because the movement of mouth muscles varies depending on the mouth shape. We deploy eight optical sensors to an HMD. The measurement points are the upper lip, upper cheek, lower lip, and cheek. These points are on both the left and right side.

We adopt two kinds of optical sensors, namely, a photo reflective sensor and position sensitive detector (PSD), which differ in sensing target and measurable range. Photo reflective sensors detect the intensity of reflected light, whereas PSDs detect the position where reflected light is received. Most photo reflective sensors can measure from about 1 mm to 20 mm, while many PSDs can measure from approximately 10 mm to 200 mm. Hence, photo reflective sensors are suitable for measuring the upper lip and the upper cheek, which are relatively close to the HMD. By contrast, PSDs are suitable for measuring the lower lip and the cheek, which are relatively far from the HMD.

We apply machine learning to recognize mouth shapes. Our system uses the multiclass classifier support vector machine (SVM) using a linear kernel, which can predict belonging probabilities to each class. Our system learns the eight optical sensor values of each mouth shape to train a classifier. This approach leads lower computational cost than processing higher dimensional data such as a camera image. Given that SVM is a supervised model, it requires the correct assignment of labels for these sensor values in the training phase. Therefore, the optical sensor values should have correct labels.

3.2. Labeling Training Data Using Vowel Recognition

We describe the relation between speech and mouth shape. Speech consists of phonemes, which are the smallest units of speech. Phonemes are characterized by the resonant frequency of air in the vocal tract. The phoneme mainly consists of consonants and vowels. Consonants are generated by dynamic movements of mouth shape, such as changes in expiratory flow or friction. Meanwhile, vowels are generated by stable movements of mouth shape, such as lip circularity and jaw opening. Focusing on such stable movements, we use vowels to label the optical sensor values, thereby enabling automatic dataset collection. Our previous study [KFJ*17], expected users to make facial expressions accord with an graphical instruction timely during the training process. On the other hand, in this study, we introduce auditory feature to label facial expression annotations to optical sensor values. However, the dataset can contain outliers if our system recognizes vowels incorrectly.

We remove outliers from the dataset. A sample in the dataset consists of eight optical sensor values. For outlier removal, we calculate the Mahalanobis distance of a sample from the mean of each class. We filter out samples higher than threshold of Mahalanobis distance from the training dataset and iterate it until all the Mahalanobis distances are less than or equal to the threshold (Formula 1).

$$X = \{X | X = \{X_0, X_1, ..., X_n\}, D_m(^{\forall}X) < \sqrt{D_{thr}}\} \quad (1)$$
$$X : SensorData$$
$$D_m(x) : MahalanobisDistance$$
$$D_{thr} : Threshold$$

We decide the optimal threshold for outlier removal. Our system obtains each training data when changing the threshold from 0.0 to 150.0 in 1.0 step. We investigate each recognition accuracy when learning each training data and find the highest value among this set of recognition accuracy. We define the proper threshold that achieves the highest recognition accuracy, thereby obtaining the best training data.

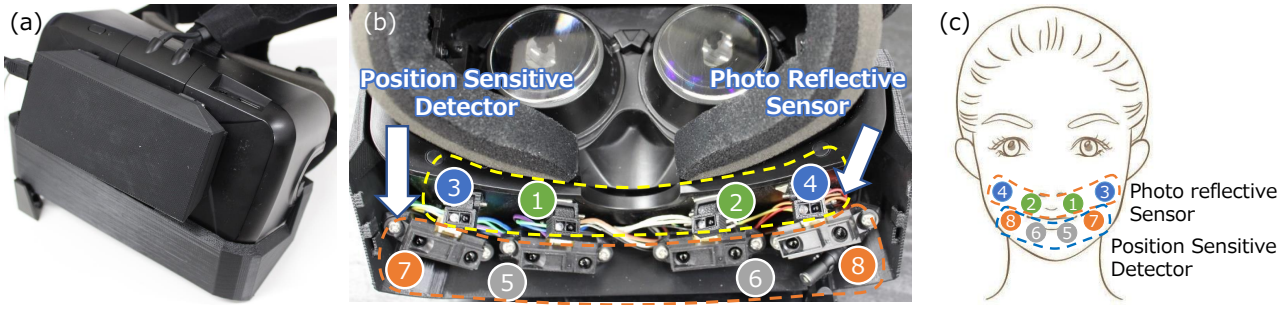

Figure 3: *(a) Prototype, (b) Placement of Sensors, (c) Point of Measurement*

3.3. Mouth Shape Reproduction

We reproduce various mouth shapes from the optical sensors. The blending of several mouth shapes is assumed to reproduce various mouth shapes. In the preparation of the parameters of several mouth shapes $\vec{P}_i$, $(i = 1, 2, \cdots, m)$, mouth shape P can be expressed as

$$\vec{P} = \vec{P}_0 + \sum_{i=1}^{m} s_i(\vec{P}_i - \vec{P}_B) \tag{2}$$

where $\vec{P}_0$ is the parameter of the mouth shape during silence, $\vec{S} = (s_1, s_2, \cdots, s_m)$ is the belonging probability to each mouth shape class.

This approach is the same as that in a previous research [KFJ*17], which synthesizes facial expressions according to five facial expressions, namely, neutral, happy, angry, surprised, and sad.

4. Implementation

Our system consisted of a computer and a device that measures mouth shape and audio. The device sent the measured sensor values and audio signals to the computer, which then learned the sensor values and recognized the mouth shapes. In training, the computer recognized vowels from the audio signals, labeled the sensor values, and used these labeled sensor values to train an SVM. In recognition, the SVM recognized mouth shapes from the optical sensor values. We describe our hardware and software in Section 4.1 and Section 4.2, respectively.

4.1. Hardware

We developed a prototype made of a modified HMD to measure the mouth shape and audio (Fig. 3(a)). The prototype had four components, namely, photoreflectors (LBR-127 HLD), optical distance measuring units (SHARP GP2Y0A21 YK), a microphone (Audio-Technica AT9904), and a microcomputer (Akitsuki Densho AE-ATMEGA 328-MINI). The photoreflectors and the optical distance measuring units were optical sensors attached to the bottom of the HMD (Oculus Rift DK2 [Ocu]). The microphone was attached to the right side of the mounting surface and connected to the computer through an amplifier (Audio-Technica AT-MA2). The audio signal was directly sent to the computer. Meanwhile, the microcomputer was attached to the front of the HMD and connected to the computer with a USB cable. The microcomputer sent the sensor values of the photoreflectors and the optical distance measuring units. Covers for the sensors and circuits were created using a 3D molding machine.

Figure 3(b) shows the arrangement of the photoreflectors and the distance measuring units. The photoreflectors (Nos. 1 to 4) measured the upper lip and the upper cheek. The optical distance measuring units (Nos. 5 to 8) measured the lower lip and the cheek (Fig. 3(c)). The detection ranges of the photoreflectors and of the optical distance measuring units were 1-10 and 100-800 mm, respectively.

4.2. Software

We implemented mouth shape recognition. Our software had two processes, namely, training and recognition. In the training process, the computer collected training data by vowel recognition. The computer recognized vowels from the audio signal and labeled optical sensor values with the recognized vowels. In the recognition process, the computer recognized the mouth shape from the optical sensor values. The computer predicted the belonging probabilities of each class and synthesized the mouth shape using these probabilities. We describe the training and recognition processes in Section 4.2.1 and Section 4.2.2, respectively.

4.2.1. Training Process

We collected the used dataset automatically by vowel recognition. The computer received eight optical sensor values and audio signals. At first, the computer recognized a vowel from the audio signals. Then, the computer labeled the eight optical sensor values with the vowel, thereby enabling the collection of the dataset of vowels. The computer provided a waiting period before such dataset is collected. During this period, the computer acquired the optical sensor values, which were labeled by the computer with "silence." Thus, the dataset was obtained.

We removed outliers from the dataset to obtain the training data. The computer calculated the Mahalanobis distance of the samples for each class in the dataset and then eliminated the samples whose Mahalanobis distances were higher than the threshold. After elimination, the training data was obtained for training the SVM.

4.2.2. Recognition Process

We recognized the mouth shapes from eight optical sensor values. The computer inputted the eight optical sensor values it received

to SVM, which then predicted the belonging probabilities to each mouth shape class. The computer detected the class that had the highest probability of these classes. The computer then regarded this class as the recognition result.

5. Experiment

We evaluated the recognition accuracy of six mouth shapes by the embedded optical sensors in Experiment 1. Five of six mouth shapes were of mouths speaking five Japanese vowels. The remaining shape was one during silence. We then collected training data manually and called the method of learning mouth shapes in this experiment "manual learning." In Experiment 2, we evaluated the recognition accuracy of mouth shapes by automatic labeling using vowel recognition. In particular, we examined how the automatic labeling method affected the recognition accuracy of the mouth shapes. We compared the recognition accuracy of six mouth shapes in Experiment 2 with the results of Experiment 1. Then, we collected training data automatically using vowel recognition and called this method of learning mouth shapes "automatic learning."

5.1. Experiment 1: Mouth Shape Recognition by Manual Learning

We investigated the recognition accuracy of the following mouth shapes by the embedded optical sensors: "silence," "a", "i", "u", "e", and "o". "Silence" had a closed mouth and no facial expression. "a", "i", "u", "e", and "o" were the mouth shapes for five Japanese vowels. The subjects of the experiment were five Japanese males in their twenties, none of whom had speech disorders or abnormalities in peripheral shapes, including the mouth. The experimental procedures were as follows.

1. We explained the six mouth shapes ("silence," "a", "i", "u", "e", and "o") to the subjects. After the explanation, we instructed the subjects to wear our prototype.
2. We instructed each subject to make the six mouth shapes and to hold them until we provided next instruction. The order of instruction was as follows: "silence," "a", "i", "u", "e", and "o". We manually operated the keyboard to collect 50 samples for each mouth shape.
3. We iterated Step 2 three times.
4. We instructed each subject to make the six mouth shapes again and to hold them until we gave additional instruction. The order of instruction was the same as in Step 2. We manually operated the keyboard to collect 200 samples for each mouth shape.

We collected 900 samples (50 samples * 6 mouth shapes * 3 iterations) for the training in Step 2. We collected 1200 samples (200 samples * 6 mouth shapes * 1 iteration) for the test in Step 4.

5.2. Result of Experiment 1

Table 1 shows the result of Experiment 1. The average recognition accuracy of five subjects was approximately 99.9%. Therefore, our method recognized the six mouth shapes with high accuracy.

Compared with a previous study on facial expression recognition [KFJ*17], our system achieved higher recognition accuracy.

Table 1: *Result of Mouth Recognition Accuracy using Optical Sensors*

Subject	A	B	C	D	E
Recognition Accuracy	100.0 %	99.3 %	100.0 %	100.0 %	100.0 %

We believe that this is because the deformation around the mouth is larger than that around the eyes. This large deformation leads to a wide variance in sensor values, which enables the accurate classification of mouth shapes.

5.3. Experiment 2: Mouth Shape Recognition by Automatic Learning

We investigated the influence of the automatic labeling method on recognition accuracy. We collected the dataset by combining vowel recognition and mouth shape measurement. Then, to obtain the best training data, this dataset was analyzed. In the analysis, we calculated the Mahalanobis Distance of sensor data in the dataset, and an optimal threshold of Mahalanobis Distance was decided for outlier removal. We used this optimal threshold to obtain the training data, which was then used to evaluate the recognition accuracy of mouth shapes. Finally, the results of this experiment were compared with those of manual learning.

5.3.1. Experiment 2a: Decision of Optimal Threshold

We investigated the optimal threshold of Mahalanobis Distance for outlier removal. We evaluated the recognition accuracy of mouth shapes under the threshold values of 1.0-150.0, with 1.0 change interval. The subjects were the same as those in Experiment 1. We then collected a dataset labeled by using vowel recognition for each subject. The following is the procedure for collecting the dataset.

1. We explained six mouth shapes ("silence," "a", "i", "u", "e", and "o") and a user interface for automatic dataset collection displayed on the HMD to each subject. We asked the subjects to speak the five Japanese vowels clearly during this experiment. Then we instructed the subjects to wear our prototype.
2. We displayed the user interface for learning mouth shapes through the HMD and used it to instruct the subjects to wait. During this time, we collected 50 samples for "silence."
3. Our system instructed the subjects to speak vowels through the user interface. Speaking instructions and gauges were displayed

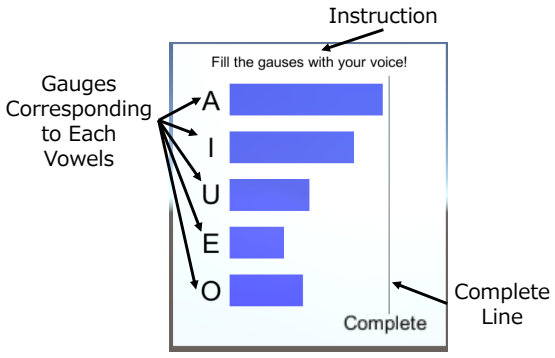

Figure 4: *Interface to Collect Dataset Automatically*

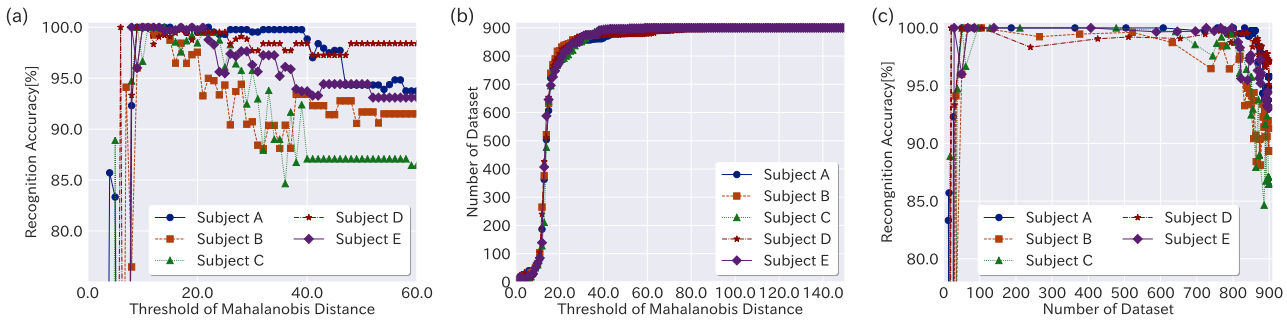

Figure 5: *(a) Variance in Classification Accuracy of Mouth Shape at Each Threshold, (b) Number of Dataset and Recognition Accuracy, (c) Threshold and Number of Dataset*

on the user interface. During this period, we gathered 50 samples for the five classes ("a", "i", "u", "e", and "o"). The order of speaking the vowels was arbitrary for the subjects.
4. Our system provided a 3 s break to the subjects by displaying a "waiting" instruction.
5. Our system iterated thrice from Step 2 to Step 4.

We collected 900 samples (50 samples * 6 mouth shapes * 3 iterations) in Step 2 and Step 3.

Figure 4 shows the interface for automatic data collection. The user interface had three components, namely, gauges, an instruction, and a completion line. The upper part of the interface displayed instructions, such as waiting and speaking. The completion line was on the right side of the interface. Arrival of the gauges at this line indicated the completion of sample collection. The center of the interface had the gauges, which indicated the number of sensor data collected for each mouth shape. As the sensor data increased, these gauges extended to the right and eventually reached the completion line. The color of these gauges indicated whether our system completed sample collection: blue meant unfinished, and red meant finished.

We analyzed the dataset to remove only outliers. In the analysis, we explored the threshold of Mahalanobis Distance to separate outliers. The following is the procedure of analyzing the dataset.

1. Our system set the threshold to 1.0.
2. Our system obtained the dataset filtered with the threshold using formula 1.
3. The filtered dataset was divided into training and test data (even and odd).
4. Our system evaluated the recognition accuracy of the threshold. Our system learned training data and calculated classification accuracy on test data.
5. If the threshold was lower than or equal to 150.0, our system added the threshold to 1.0 and return to Step 1. If not, our system finished the analysis.

Thus, we gathered the recognition accuracy corresponded to each thresholds, which was 1.0-150.0, with 1.0 change interval. Among this set of recognition accuracy, our system detected the threshold which achieved the highest recognition accuracy.

5.3.2. Result of Experiment 2a

Figure 5(a) shows the experiment results where the recognition accuracy was 80%-100% and the threshold was 0.0-60.0. The recognition accuracy for classes without samples was 0.0%, as our system could not recognize the mouth shape. We found that several thresholds approximately between 10.0 and 25.0 achieved the highest accuracy (Table 2).

Table 2: *Maximum Recognition Accuracy and Thresholds of Each Subjects*

Subject	Highest Recognition Accuracy	Threshold
A	100.0 %	9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 17.0, 19.0, 22.0
B	100.0 %	10.0, 11.0
C	100.0 %	8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0
D	100.0 %	6.0, 9.0, 10.0, 11.0
E	100.0 %	8.0, 10.0, 11.0, 12.0, 13.0, 17.0, 19.0, 21.0

Figure 5(b) shows the graph of the threshold and the number of datasets after outlier removal. Figure 5(c) is the graph of the number of datasets after outlier removal where the recognition accuracy was 80%-100%.

Figure 5(b) reveals that the number of datasets significantly fluctuated between approximately 100 and about 800 when the threshold was between 10 and 25. According to Fig. 5(c), the recognition accuracy decreased when the number of the dataset was out of the 100-700 range. This finding implies that the shortage of the dataset and insufficient removal of outliers resulted in the decreased recognition accuracy. Therefore, we considered that the optimal threshold was the maximum value among Table 2. For example, the optimal threshold for subject A was 22.0.

5.3.3. Experiment 2b: Mouth Shape Recognition in Automatic Learning

We evaluated the recognition accuracy by the automated labeling method by comparing the recognition accuracy in this experiment with those of Experiment 1. We conducted the automatic learning

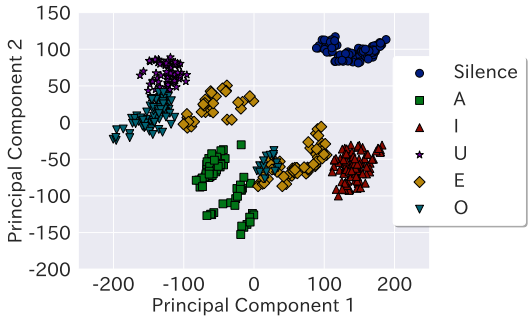

Figure 6: *PCA Result of Subject C's Training Data*

experiment after Experiment 2a without removing the HMD, and the subjects were the same as those in Experiment 1.

We removed outliers from the dataset to obtain the training data. We used the maximum threshold in Table 2 and collected 1,200 samples (200 samples * 6 mouth shapes * 1 iteration) to test each subject. The procedure of data collection was the same as in Step 4 of Experiment 1.

5.3.4. Result of Experiment 2b

Table 3 shows the result of Experiment 2b. The average recognition accuracy of the five subjects was approximately 96.3%. Compared with the recognition accuracy in manual learning, that in automatic learning decreased by about 3.6%. Nonetheless, our automatic labeling method classifies the six mouth shapes accurately.

Table 3: *Result of Mouth Recognition Accuracy using Optical Sensors in Automated Labeling Condition*

Subject	A	B	C	D	E
Recognition Accuracy	100.0 %	95.7 %	86.1 %	100.0 %	100.0 %

The recognition accuracy of subject C decreased significantly compared with that in Experiment 1. For analysis of this result, we visualized subject C's training data with principal component analysis (PCA) in Fig. 6. We also showed subject C's confusion matrix of mouth shape recognition in Table 4. According to Fig. 6, some samples of "e" were close to clusters of "i" and "o". In collecting the dataset, we thought that his mouth shape of "e" differed at each trial. Therefore, outliers of subject C's "e" increased. This large number of outliers led to an increase in the Mahalanobis distances of all samples and thus insufficient outlier removal. Table 4 indicated that our system mispredicted 53.5% of "i" as "e" and mispredicted 30.0% of "e" as "o". From these findings, we found that our mouth shape recognition accuracy depended on the stability of the reproduction of the user's mouth shape.

6. Application: Reflecting Mouth Shape on Avatar

We developed an application that transferred the HMD user's mouth shape to an avatar (Fig. 7). This application blended the parameter of the six mouth shapes ("silence," "a", "i", "u", "e", and "o") to reproduce the mouth shapes.

We described the procedure by which this application reflected

Table 4: *Subject C's Confusion Matrix of Mouth Shape Recognition*

Correct Label \ Predicted Label	Silence	A	I	U	E	O
Silence	1.00	0.00	0.00	0.00	0.00	0.00
A	0.00	1.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.47	0.00	0.54	0.00
U	0.00	0.00	0.00	1.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.70	0.30
O	0.00	0.00	0.00	0.00	0.00	1.00

the mouth shape on the avatar. Our system predicted the probabilities of each mouth shape from the optical sensor values. By using these probabilities, the application calculated the parameter of the mouth shape blended on the basis of Formula 2 and then applied this parameter to the avatar. Figure 8 shows that the avatar reflected the blended mouth shape. This technique enabled the reflection of the various movements around the mouth, which were not limited to six states.

7. Discussion

Our system recognized the mouth shape during silence and during the speaking of the five Japanese vowels. However, our system encountered difficulties in reproducing certain mouth shapes (e.g., lip biting, cheek swelling). In future work, we explore the base of the mouth shape to potentially represent complex mouth shapes.

Although we reproduced various mouth shapes by blending six mouth shapes, we did not discuss the suitability of using the belonging probabilities as weights of blending. Therefore, the system may not have blended the mouth shape accurately. In the future, we plan to investigate the geometry deformation of mouth shapes to explore a suitable blending method.

Our current system has a training process for each individual to build a mouth shape classifier. If we can build a universal classifier that estimate any user's mouth shapes, this process can be omitted.

We used only optical sensor values for recognizing mouth shapes after the training process. We may be able to use auditory infor-

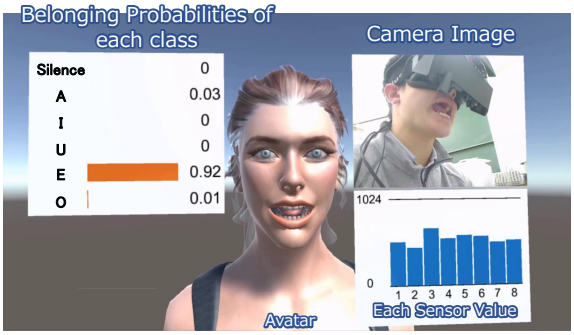

Figure 7: *Application Reflecting User's Mouth Shape on Avatar*

Figure 8: *Reflecting Animation of User's Various Mouth Shape to Avatar*

mation as an additional feature to recognize mouth shapes more robustly.

8. Conclusion

We propose a system that recognizes the mouth shapes of HMD users with optical sensors. We developed an HMD-based prototype equipped with four photoreflectors, four optical distance measuring units, a microphone, and a microcomputer. This prototype measured mouth shape and audio signals. The photoreflectors and optical distance measuring units detected the movement of eight points (upper lip, upper cheek, lower lip, and cheek). The microphone acquired audio signals. Meanwhile, our system detected the vowels from audio and used them to label the optical sensor values.

From the manual learning experiment, our system achieved an average accuracy of approximately 99.9% for the five subjects. We also evaluated the recognition accuracy by the automatic labeling method, which achieved an average accuracy of about 96.3% for all subjects. The recognition accuracy of automatic learning was lower by about 3.6% than that of manual learning. Nonetheless, we believe that our system could label training data properly through our experiments.

We also developed an application that projected the mouth shapes to an avatar. The application predicted the belonging probabilities to each mouth shape class, and we blended each mouth shape on the basis of the belonging probabilities to reproduce various mouth shapes. This application showed that our system could reflect various mouth shapes on the avatar.

Acknowledgements

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

References

[GP05] GORANKA Z., PANDZIC I. S.: A real-time lip sync system using a genetic algorithm for automatic neural network configuration. In *2005 IEEE International Conference on Multimedia and Expo* (July 2005), pp. 1366–1369. doi:10.1109/ICME.2005.1521684. 2

[GS14] GRUEBLER A., SUZUKI K.: Design of a wearable device for reading positive expressions from facial emg signals. *IEEE Transactions on Affective Computing 5*, 3 (July 2014), 227–237. doi:10.1109/TAFFC.2014.2313557. 2

[HDS*19] HICKSON S., DUFOUR N., SUD A., KWATRA V., ESSA I.: Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Jan 2019), pp. 1626–1635. doi:10.1109/WACV.2019.00178. 2

[IYY01] ITOI K., YASUSHI M., YUKIO K.: Intelligent coding of facial expression using neural network and morphing. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)* (May 2001), pp. 352–355. doi:10.1109/ISIMP.2001.925406. 2

[KFJ*17] KATSUHIRO S., FUMIHIKO N., JIU O., KATSUTOSHI M., YUTA I., YUTA S., MAKI S.: Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)* (March 2017), pp. 177–185. doi:10.1109/VR.2017.7892245. 1, 2, 3, 4, 5

[LTO*15] LI H., TRUTOIU L., OLSZEWSKI K., WEI L., TRUTNA T., HSIEH P.-L., NICHOLLS A., MA C.: Facial performance sensing head-mounted display. *ACM Trans. Graph. 34*, 4 (July 2015), 47:1–47:9. URL: http://doi.acm.org/10.1145/2766939, doi:10.1145/2766939. 2

[MSO*16] MASAI K., SUGIURA Y., OGATA M., KUNZE K., INAMI M., SUGIMOTO M.: 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* (New York, NY, USA, 2016), IUI '16, ACM, pp. 317–326. URL: http://doi.acm.org/10.1145/2856767.2856770, doi:10.1145/2856767.2856770. 2, 3

[Ocu] OCULUS: https://www.oculus.com/. 4

[OLSL16] OLSZEWSKI K., LIM J. J., SAITO S., LI H.: High-fidelity facial and speech animation for vr hmds. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 221:1–221:14. URL: http://doi.acm.org/10.1145/2980179.2980252, doi:10.1145/2980179.2980252. 2

[SMK*17] SAKASHITA M., MINAGAWA T., KOIKE A., SUZUKI I., KAWAHARA K., OCHIAI Y.: You as a puppet: Evaluation of telepresence user interface for puppetry. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2017), UIST '17, ACM, pp. 217–228. URL: http://doi.acm.org/10.1145/3126594.3126608, doi:10.1145/3126594.3126608. 2

[Uni] UNITY O. L.: https://developer.oculus.com/downloads/package/oculus-lipsync-unity/. 2

[YMM07] YUKI M., MADDUX W. W., MASUDA T.: Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology 43*, 2 (Mar. 2007), 303–311. doi:10.1016/j.jesp.2006.02.004. 1