International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2017)
R. Lindeman, G. Bruder, and D. Iwai (Editors)

Facial Performance Capture by Embedded Photo Reflective Sensors on A Smart Eyewear

Nao Asano, Katsutoshi Masai, Yuta Sugiura, Maki Sugimoto

Keio Univerisity, Japan

Abstract

Facial performance capture is used for animation production that projects a performer's facial expression to a computer graphics model. Retro-reflective markers and cameras are widely used for the performance capture. To capture expressions, we need to place markers on the performer's face and calibrate the intrinsic and extrinsic parameters of cameras in advance. However, the measurable space is limited to the calibrated area. In this paper, we propose a system to capture facial performance using a smart eyewear with photo reflective sensors and machine learning technique.

CCS Concepts

•***Hardware*** → *Sensor devices and platforms;*

1. Introduction

One's facial expression conveys non-verbal information, such as one's emotional state and psychological state, to others. It plays a vital role in conveying meanings that cannot be expressed in words in the context of everyday communication and content creation, such as sitcoms and human drama. In the field of computer graphics (CG), facial performance capture systems are used to acquire the facial expressions of performers continuously. By reflecting the facial expressions to a three-dimensional (3D) face model, it is possible to produce an animation without performing a significant amount of physical calculation.

Common facial performance capture techniques use retro-reflective markers and multiple cameras to capture the expressions robustly [Wil90, BGY*13, HCTW11]. However, such systems need many markers on facial skin for every recording. In addition, they require the pre-calibration of internal and external camera parameters, such as lens distortions and the position and orientation of the cameras, and we cannot move the cameras during capturing. Other methods use monocular cameras [GVWT13, WBGB16, CBZB15, RHKK11] and commodity RGB-D cameras [BWP13, WBLP11] to reduce the cost of pre-calibration. These methods capture faces with a small hardware setting, but the space that can track a face is still limited due to the viewing angle of the camera and occlusions.

There are several methods for recognizing facial expressions using wearable devices instead of cameras [GS14, SFP99, MSO*16]. Masai et al. developed a system to classify basic facial expressions using a smart eyewear with photo-reflective sensors and machine learning [MSO*16]. They measure the proximity between the skin surface on a face and the eyewear with the sensors. Their method allows the classification of facial expressions in many situations.

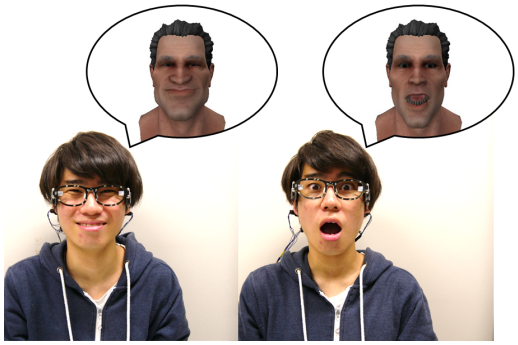


Figure 1: *Our system estimates 3D points on the skin surface using smart eyewear. We can retarget the facial expression to the 3D model from the points.*

However, the authors do not try to estimate geometry change with different facial expressions.

In this research, we propose a facial performance capture system using the smart eyewear developed by Masai [MSO*16]. We combine the external device also to obtain the information of the movement of the mouth. Then, we construct a regression model that expresses the relationship between the marker position of the motion capture system and values of the photo reflective sensors and estimate the 3D position of the skin surface from the wearable device using the regression model. We evaluate the estimation accuracy of the marker position. We show that our system can capture facial performance by applying the estimated marker position to 3D face model (Figure 1).

DOI: 10.2312/egve.20171334

2. Related Works

2.1. Facial performance capture using cameras

In the field of facial performance capture, marker-based techniques are widely used in research and content production scenarios [Wil90, BGY*13, HCTW11]. Williams developed a method to control the face of a computer-generated animation model [Wil90]. His method follows the facial expressions of multiple retro-reflective markers affixed to the skin surface of the user's face. Commercial systems such as Natural Point's Optitrack or Vicon acquire facial expressions by tracking the light reflected by the marker with multiple cameras and measuring the 3D position in real space. Although this method can accurately estimate the 3D position, there are three problems. First, the preparation cost is high, such as the arrangement of markers and calibration of cameras installed in the environment. Second, the processing cost is also high because a large number of camera images are used. Besides, the measurable space is limited to the calibrated area.

Marker-less methods have also been studied to reduce the cost of pre-calibration. Structured light techniques can compute the depth map of the moving facial surface. The techniques use the projected light pattern texture measured with a camera [PVGO96, WLVGP09]. Zhang and colleagues generated an individual face model offline [ZSCS04]. The authors fitted a face template mesh to depth map and combined it with optical flow. Passive capture (multi-view capture) using multiple two-dimensional (2D) images have also been studied [BHPS10, GFT*11, BHB*11]. Ghosh and colleagues statically reconstructed facial geometry from diffuse and specular reflectance information using polarized spherical gradient illumination [GFT*11]. Also, Beeler et al. dynamically reconstructed the pore-level geometry including wrinkles and folds from seven camera images by considering the similarity of facial expressions between different frames [BHB*11]. Compared with the marker-based techniques, a marker-less method can acquire spatiotemporal geometry of the face with a higher resolution. However, it still requires the preparation cost to set projectors, special light sources, multiple cameras, etc.

To reduce the complexity of hardware setting in multi-view capture, researchers investigated the methods of densely reconstructing geometric shapes with a pair of stereo cameras [VWB*12] and monocular cameras [GVWT13, WBGB16]. Wu et al. reconstructed facial performances from a monocular camera using a local deformation model with anatomical constraints [WBGB16].

The researchers developed the methods by lowering the resolution to reduce the computational cost and to enable real-time facial performance capture and retargeting. These methods use commodity RGB-D cameras [BWP13, WBLP11] or monocular cameras [CBZB15, RHKK11]. Weise et al. calculated the blend shape parameter of the 3D face model from the commodity RGB-D camera, so that the user can control the facial expressions of animations in real-time at a low cost [WBLP11]. In these methods, there are relatively few instruments used for measurement, and preparations in advance are simplified. However, the limitation of tracking space using a camera is still a problem.

2.2. Facial performance capture using sensors

Research has been done to acquire the facial expressions of users without cameras. Speech-driven animation, a technique that controls face models based on speech signals, can reproduce facial expressions based on language information [Bra99, CB05]. However, the user needs to say something when acquiring facial expressions. Moreover, it is difficult to obtain the movements with low relevance to speaking, such as those of the eyebrows and eyes.

Several studies have used a contact-type sensor for acquiring facial expressions. Scheirer et al. recognized the movement of facial muscles related to facial expressions such as confusion and interest using a piezoelectric sensor [SFP99]. Li et al. placed the strain sensor inside the HMD and estimated the shape of the upper part of the face covered with the HMD from the sensor values. They combined this with the RGB-D camera installed in the HMD to acquire the expression of the whole face [LTO*15]. Jorge et al. controlled the mesh model of the face using the EMG sensors [LM99]. The EMG signal from the face is useful information for acquiring the movement of the muscles of the face. The authors put the electrodes on a set of modeled facial muscles. Gruebler and Suzuki developed a wearable device that can recognize positive facial expressions using EMG sensors [GS14]. Although these methods do not use cameras in environments to acquire facial expressions, it is necessary for the sensors to be in close contact with the skin surface of the face. This feature causes concern about the stability of the device position and the comfortability when we wear it for a long time.

On the other hand, photo reflective sensors can measure facial expressions without contact [FTT13, MSOI13, MSO*16]. Masai et al. classified basic facial expressions of users using eyewear-type devices with multiple infrared photo reflective sensors [MSO*16]. Since the distance to the skin surface from the sensor changes with different facial expressions, the reflection intensity information of the sensor also changes. This device can be used in daily life. However, the authors do not try to estimate geometric changes with different facial expressions.

3. Proposed Method

In this study, we present a facial performance capture system using a smart eyewear where the photo reflective sensors are embedded. We estimate facial geometry from the reflection intensity information acquired by the sensors in the device. This method is divided into a training phase and an estimation phase (Figure 2). In the training phase, we acquire the dataset from the motion capture system and the eyewear device. Concretely, we obtain the position of the markers used for motion capture correlating to facial geometry from the motion capture and reflection intensity information from the eyewear. We convert the intensity information into the distance and reduce the dimension of the marker position by applying principal component analysis (PCA). Then, we generate a regression model representing the relationship between the two. In the estimation phase, using the regression model, we estimate the 3D position of the markers representing the facial expression shape of the user from the sensor information acquired by the device. We retarget the facial expression to the 3D model based on the estimated marker position.

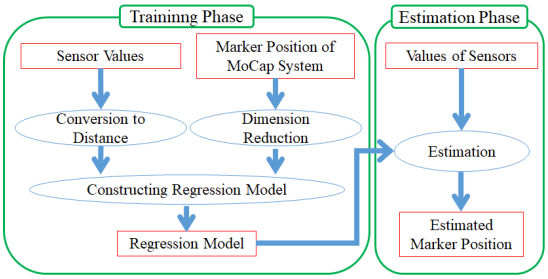


Figure 2: *Our system constructs a regression model that expresses the relationship between the marker position and values of the sensors in training phase and uses the regression model for estimation.*

3.1. Acquisition of mouth opening and closing using photo reflective sensors

The eyewear developed in [MSO*16] cannot measure the deformation of the lower jaw, etc., where the change does not appear on the cheek. In this research, we measure the skin deformation of the whole face. Therefore, we made the external device to acquire the information related to the approximate shape of the mouth. We combined the eyewear device with the external device. The sensors on the external device measure the temporomandibular joint and lower jaw where skin deformation occurs with the opening and closing of the mouth.

3.2. Conversion from reflection intensity information to distance information

According to [MSO*16], reflection intensity information acquired by the sensors has a nonlinear relationship with the distance to the skin surface. Since it is expected that the distance can be approximated to the linear relationship with the marker position information, the conversion from reflection intensity information to the distance value by polynomial approximation prevents the regression model from becoming complicated.

3.3. Dimension reduction of marker position information using PCA

Since the number of dimensions of the sensors information is smaller than that of the markers position information, it is difficult to construct a regression model directly. Therefore, we perform the dimension reduction process of the marker position using PCA. The dimension reduction by PCA is used to construct a linear model expressing a non-rigid surface [SPIF07] or face [BV99, LCXS09] as a linear combination of basic shapes, and it is possible to represent arbitrary shapes with fewer parameters. We apply PCA to the marker position information in the training dataset to generate a model that expresses the position of the markers on the face with few parameters.

The positions of the V markers are expressed as vectors of $3V$ dimension since each marker is represented by 3D coordinates. By applying PCA to the space of $3V$ dimension where the marker position information of the training data is distributed We retain N_c principal components $\boldsymbol{S_k} = \{s_{k,1}, \ldots, s_{k,3V}\} \in \mathbb{R}^{3V} (1 \leq k \leq N_c)$.

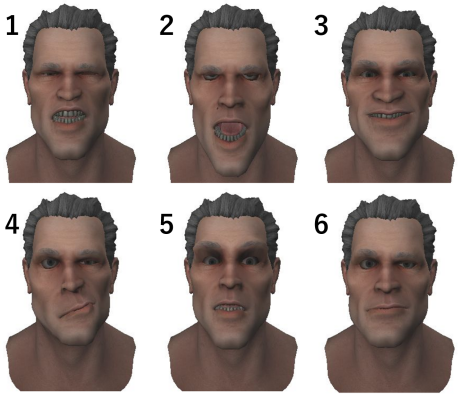


Figure 3: *Each principal component shows the representative facial expression change of the user. The principal component score corresponds to the weight of the expression.*

Vector $\boldsymbol{S}$ corresponding to an arbitrary marker arrangement can be written as Equation 1. The principal component scores are ω_k, and the average value of each marker position is stored in the array $\bar{\boldsymbol{S}}$. We obtain the principal component scores by the regression model (Equation 2) in Section 3.4.

$$\boldsymbol{S} = \bar{\boldsymbol{S}} + \sum_{k=1}^{N_c} \omega_k \boldsymbol{S_k} \tag{1}$$

The principal components represent the position changes of all markers corresponding to the user's representative facial expressions. We can represent the arbitrary marker arrangement by the linear combination of principal components. Therefore, using this method, we can represent the marker position of the $3V$ dimension of the skin surface at the time of arbitrary expression using the N_c dimensional principal component scores.

3.4. Constructing the Regression Model

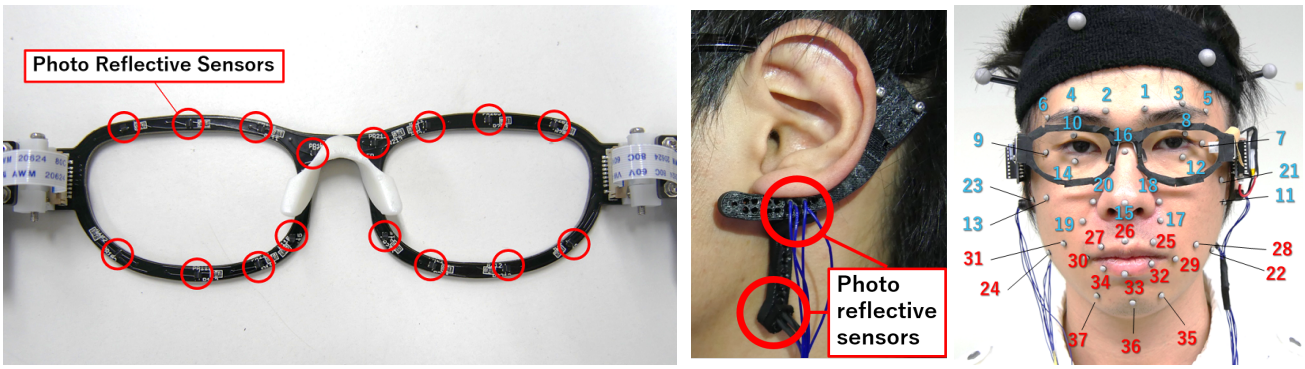
To estimate the marker displacement using the sensors of the device, we construct a linear regression model that expresses the relationship between sensor information and each principal component score.

We make N_c regression models with explanatory variables as D photo reflective sensor information and object variables as principal component scores of the k th principal component. When the photo reflective sensor information is x, the coefficients are w_k, ω_k corresponding to the principal component scores can be expressed as follows:

$$\omega_k = w_0 + w_1 x_1 + \ldots + w_D x_D \tag{2}$$

3.5. Estimation of the Marker Position

In the estimation phase, we predict the principal component scores from the sensor information. The sensor information is converted into the distance information. We predict with the regression model generated in Section 3.4. After that, as described in Section 3.3, the marker positions are estimated by the principal component scores.



(a) (b) (c)

Figure 4: *Our system settings (a) sensor layout of eyewear; (b) the extension device; (c) layout of markers*

It can reflect in the 3D face model based on the estimated position information of the marker. In addition, the proposed method does not require a motion capture system after constructing the regression model and can estimate the 3D point of the face surface without being subject to spatial constraints.

4. Implementation

We extended the eyewear arranged with 16 photo reflective sensors around the eyes (Figure 4a). The extension part is an ear-hook type. The two parts are attached to both sides of the temples. We created the body of the extension parts with a 3D printer (Figure 4b). We placed the same photo reflective sensors (Kodenshi-SG 105) as those used in [MSO*16] near the temporomandibular joints and lower jaws and added four sensors in total. The sensors on the extension part were wired to Arduino Uno. We used serial communication with client applications at a speed of 250,000 bits per second. We attempted to reduce the influence of ambient light by switching the sensor LED on and off and measuring the difference, as shown in [MSO*16].

For the motion capture system software, we used Natural Point's Arena Expression Facial Motion Capture 1.8.6. The motion capture system software was operated on the desktop computer. UDP communication was performed with the desktop computer as the server and the notebook computer as the client. The data was transmitted to the client application. If the device occludes some markers, the system cannot calculate the positions of them. Therefore, we arranged the markers so that the cameras can capture them regardless of the facial expression change (Figure 4c). Moreover, when photo reflective sensors receive infrared light emitted by the LED incorporated in the camera of the motion capture system, the sensor value is much affected. Therefore, we use only the sensor values which are when the strobe LED does not emit light by the median filter taking the median value of 10 consecutive samples.

5. Evaluation

5.1. Evaluation Procedure

To verify the proposed method, we calculated the estimation error of the marker position. We had three participants in this experiment. In the training phase, we asked each participant to wear the device. Then, we placed the markers on the face. We asked the participants to make the same facial expression as the expression shown on the screen. The dataset of the expressions was pre-made. We switched the expressions every 3 seconds. While the user made various expressions, we acquired the sensor information of the device and the marker position information of the MoCap system at the same time. We acquired a total of 5,000 datasets from each participant.

We presented about 600 expressions to the participants. These facial expressions are based on the first to sixth principal components of the marker position information measured by motion capture before the experiment. We randomly chose each value of the components to the maximum, the origin, and the minimum, then applied to the 3D model. We excluded outliers among the acquired datasets. For outlier rejection, we set the value at four times of the standard deviation of the marker positions. The number of outliers for each participant was 165, 411, and 457 respectively. The number of samples differed among the expressions since the marker could not follow up and the measurement temporarily stopped. The training phase took about 30 minutes.

In the estimation phase, we created a total of 19 facial expressions: a neutral expression, (A) six facial expressions in which values of the first to six principal components are randomly set as long as the expression can be made by everyone, (B) six universal facial

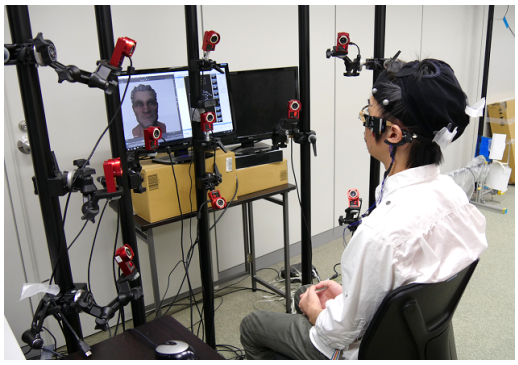


Figure 5: *Experiment environment*

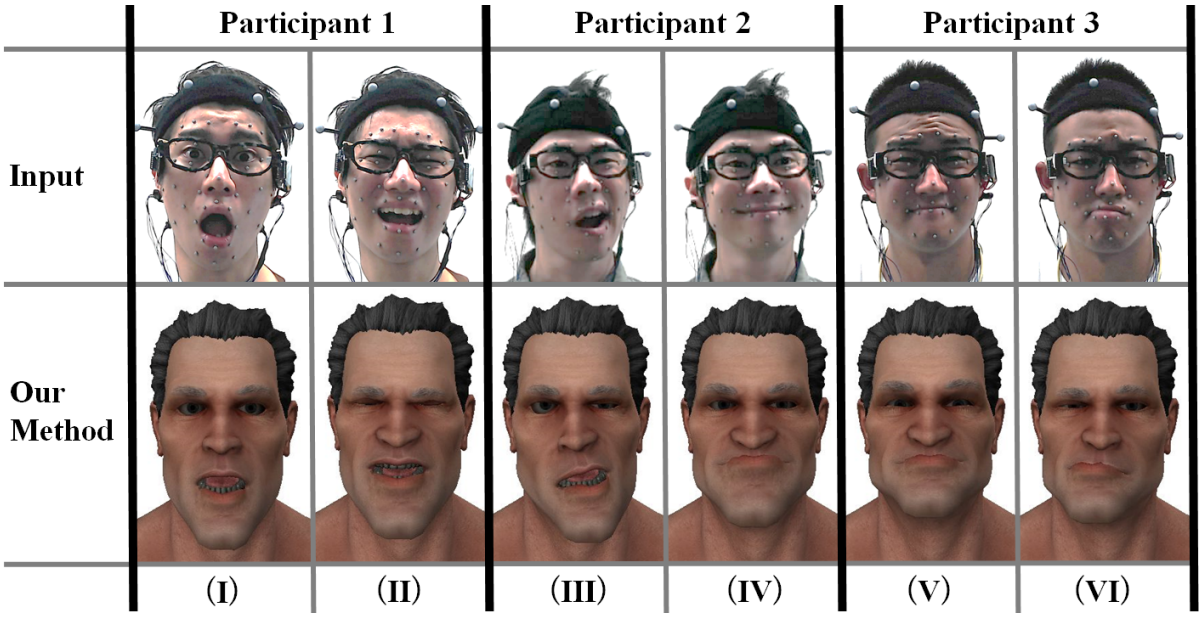


Figure 6: *The user face and the results of retargeting the marker position information estimated by the proposed method*

expressions (happiness, disgust, anger, surprise, fear, and sadness) defined by Ekman [EF71], (C) asymmetric expressions that are not included in the basic facial expressions. In the same way as the training, we presented the facial expressions for 10 seconds each, and we let them make the same facial expression. At this time, the estimated value is the marker position information estimated from the sensor while the true value is the marker position information measured by the motion capture system. The estimation error is the root mean squared error for each marker. We considered the first 20 principal components for the regression model. The cumulative contribution rate up to the 20th principal component was 99.0%.

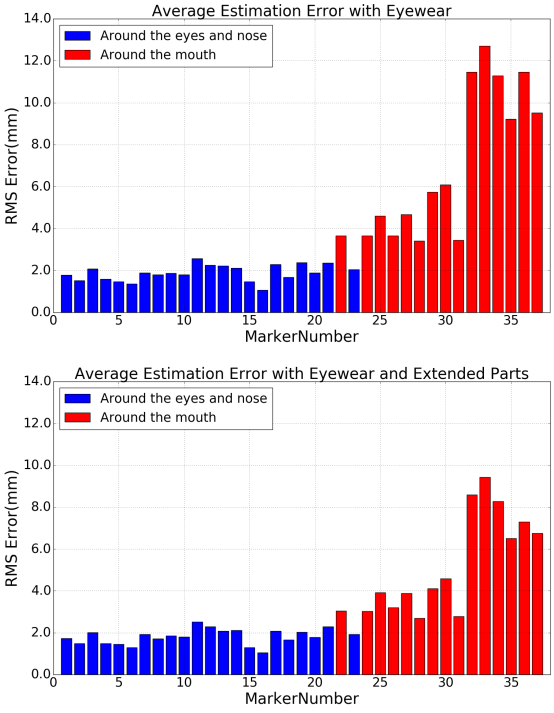


Figure 7: *Average estimation error of all participants with 16 sensors on eyewear and with 20 sensors including the extended parts*

5.2. Results

Figure 6 shows the comparison of facial expression retargeting between the user's face as a ground truth and the result of retargeting the marker position information estimated by the proposed method to the 3D model. By comparing the user face with the estimated expression, our method can estimate the eyebrows, the movement of the jaws in (I), (II), (III), and the pulling up of the cheeks in (IV), (V). The system estimated the mouth shape based on the cheek deformation and jaw joint movement. This made it difficult to estimate the precise shape of the lips. We only can estimate with less accuracy the raising of the mouth corner, as shown in (IV), and the expression to deform the lower lip, as shown in (VI). Also, the estimation is insufficient for the movement of the eyelids and eyes in (V). We assumed that it was difficult to estimate the positions of these parts because the sensor values do not much change at the time of facial expression change.

For the verification, we estimated the average values of estimation errors for each marker of three subjects. Figure 7 shows the result of the estimation using only 16 sensors around the eyes and the result of the estimation with 20 sensors by adding the extended device. The mean error for each type of expression is (A) 3.24 mm, (B) 3.25 mm, and (C) 3.16 mm. The average errors for each participant were 3.76 mm, 2.77 mm, and 3.05 mm. When using 16

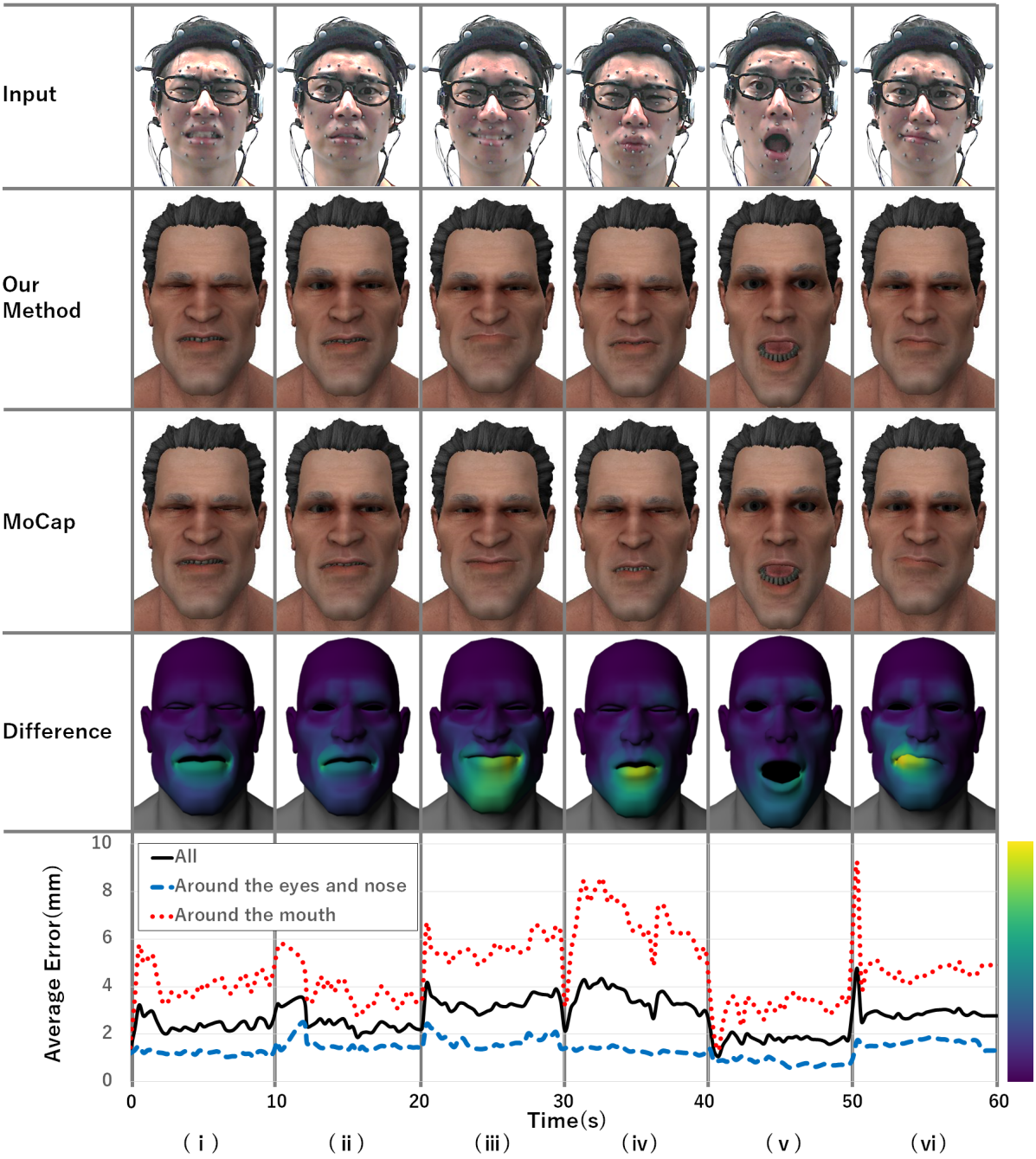


Figure 8: *Time transition of the average error of markers and the heat map showing the difference between the result of our method and that of the MoCap system. (i) disgust, (ii) fear, (iii) happiness, (iv) sadness, (v) surprise, (vi) one of the training sets.*

sensors, the average error of all subjects was 3.95 mm. The average error of the eyes and the nose was 1.89 mm, and the average error around the mouth was 6.97 mm. When using 20 sensors, the average error of all subjects was 3.19 mm. The average error of the eyes and the nose was 1.81 mm, and the average error around the mouth was 5.21 mm.

5.3. Discussion

The average error of the markers around the eyes and nose was remarkable even with 16 sensors (1.89mm). Although the accuracy of estimation varies depending on the number of sensors near the markers, Figure 7 shows that the extended parts reduced the error around the mouth. The device added in this study improved the

accuracy of estimating the shape around the mouth 1.76 mm in average. However, still, we can see the estimation errors around the jaw and lower lip, even after including the extended device. The error of the lower lip (marker numbers 32 to 34) was larger than that of the jaw (marker numbers 35 to 37). Adding more sensor around mouth improves the estimation result.

Figure 8 shows a time transition of the average error of all markers when estimating the marker position of one subject by the proposed method with 10,000 samples. The heat map showing the error distribution of each vertex when retargeting to the 3D model. As shown in the figure, the error around the eyes, nose, and the cheek are low over different expressions, and the operation is stable even when the facial expression changes. However, the high error appears in the lower lip through-out the entire series, especially when an expression that deforms only the lower lip like (iv) occurs (the estimation error exceeds 8 mm). The reason for this is a small deformation of the lower lip hardly appears in the cheeks and jaw joints on which the photo reflective sensors are arranged. Furthermore, the facial expressions shown to the subjects in the training phase were based on only the first to sixth principal components. Since these principal components with a high contribution ratio show typical deformation (Figure 3), we consider adding more training expressions with minor deformation to improve the accuracy. For the jaw, the error in (v) is not particularly significant, so there is a possibility of estimating the opening and closing. However, when the chin rises, as shown in (iii), or when the force is put in the mouth, and the jaw is displaced forward, as shown in (iv), the error is larger than the opening and closing. In this paper, the sensors were arranged mainly to acquire the information of the opening and closing of the mouth. The movement of the jaw when the high error comes out and the opening and closing of the mouth are different in places where the skin deformation occurs. Therefore, we can solve the problem by increasing the sensor in the extended device.

Since the marker-based motion capture systems have high accuracy in the measurement of feature points on the face surface, we used it for evaluating the accuracy by the device. However, the system sometimes incorrectly labeled the marker. The reason for this is that the tracking algorithm of the motion capture system is not supposed to be used with the device. Therefore, when the device occluded a part of the marker, a lot of irregular errors occurred in the measurement of markers positions. In this research, we attempted to exclude them as outliers, but even if labeling is done correctly, the measurement accuracy of the marker is lowered by the occlusion. Furthermore, the motion capture system measures the markers positions with infrared light. Since the sensors on the device measure the reflection intensity of the infrared light, the infrared light from both can interfere with each other. The interference causes noises on the sensors. We reduced the noise with the median filter. However, this filtering process lowers the data acquisition speed and takes longer time for the training phase. Considering the preparation cost and the practicality, the RGB cameras with little interference can be the alternative. Although the accuracy gets lower than that of marker-based techniques, it simplifies the system since the user do not need to attach any marker on the face.

6. Limitations and future work

The system requires each user to learn and generate a regression model. The eyewear system measures the distances from the sensors on the device to the face. However, since the distance to the face is different for each user, it is difficult for users to share the same regression model. Also, the position of the device can be shifted during use or when remounting it. These may require users to re-calibrate and generate a new model. We think that this problem can be solved by applying the normalization method proposed in [MSO*16] based on each sensor value when a user makes the neutral expression and the range of each sensor value when moving facial muscles.

We want to make applications based on the proposed system in the future. The first scenario is for live performance. This system is suitable for seeing voice actors and their representative animated characters on the stage because it is not necessary to place retro-reflective markers on actors' face surface after training. By retargeting the facial expression of an actor on the stage to an animated character, we can make a new form of live performance in which simultaneously both the actor and the animated character perform a show. Since the space for tracking the face is not limited as in the camera-based method, it is possible to follow the expression of the actor even if the movement range of the actor is broad. The second scenario is a new communication system. By using the smart eyewear with a small display, the users can conduct a conversation via an avatar that reflects the expression of the user. This system is useful as a social networking service because users can communicate non-verbal information using their facial expressions without showing their faces. Users can communicate through the system in daily situations thanks to the social acceptability of the device.

7. Conclusions

In this study, we proposed a facial performance capture system using photo reflective sensors placed on smart eyewear. Our performance capture system estimates 3D points on the skin surface using a regression model. As the implementation, we designed an extended sensor unit of the smart eyewear placed around ears to improve the accuracy of our capture system in the mouth area. To make the regression model, we obtained training datasets with a conventional motion capture system and the photo reflective sensors and applied machine learning techniques. In the evaluation, we experimented and measured the estimation error of the marker position to validate the accuracy of our system. As a result of the experiment, the average total error of all markers was 3.19 mm. In detail, the average error of the position of the markers around the eyes and nose was 1.81 mm, and that of around the mouth was 5.21 mm with the extended sensor unit. The experiment confirmed that the estimation accuracy of the marker around the mouth improved 1.76 mm by integrating the sensors.

Acknowledgements

This research was partially supported by JST CREST (JPMJCR14E1) and JSPS KAKENHI (16H05870).

References

[BGY*13] BHAT K. S., GOLDENTHAL R., YE Y., MALLET R., KOPERWAS M.: High fidelity facial animation capture and retargeting with contours. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2013), SCA '13, ACM, pp. 7–14. i, ii

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 Papers* (New York, NY, USA, 2011), SIGGRAPH '11, ACM, pp. 75:1–75:10. ii

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 Papers* (New York, NY, USA, 2010), SIGGRAPH '10, ACM, pp. 41:1–41:10. ii

[Bra99] BRAND M.: Voice puppetry. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1999), SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., pp. 21–28. ii

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1999), SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., pp. 187–194. iii

[BWP13] BOUAZIZ S., WANG Y., PAULY M.: Online modeling for realtime facial animation. *ACM Trans. Graph. 32*, 4 (July 2013), 40:1–40:10. i, ii

[CB05] CHUANG E., BREGLER C.: Mood swings: Expressive speech animation. *ACM Trans. Graph. 24*, 2 (Apr. 2005), 331–347. ii

[CBZB15] CAO C., BRADLEY D., ZHOU K., BEELER T.: Real-time high-fidelity facial performance capture. *ACM Trans. Graph. 34*, 4 (July 2015), 46:1–46:9. i, ii

[EF71] EKMAN P., FRIESEN W. V.: Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology 17*, 2 (1971), 124–129. v

[FTT13] FUKUMOTO K., TERADA T., TSUKAMOTO M.: A smile/laughter recognition mechanism for smile-based life logging. In *Proceedings of the 4th Augmented Human International Conference* (New York, NY, USA, 2013), AH '13, ACM, pp. 213–220. ii

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph. 30*, 6 (Dec. 2011), 129:1–129:10. ii

[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. i, ii

[GVWT13] GARRIDO P., VALGAERT L., WU C., THEOBALT C.: Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph. 32*, 6 (Nov. 2013), 158:1–158:10. i, ii

[HCTW11] HUANG H., CHAI J., TONG X., WU H.-T.: Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. *ACM Trans. Graph. 30*, 4 (July 2011), 74:1–74:10. i, ii

[LCXS09] LAU M., CHAI J., XU Y.-Q., SHUM H.-Y.: Face poser: Interactive modeling of 3d facial expressions using facial priors. *ACM Trans. Graph. 29*, 1 (Dec. 2009), 3:1–3:17. iii

[LM99] LUCERO J. C., MUNHALL K. G.: A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America 106*, 5 (1999), 2834–2842. ii

[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. ii

[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. i, ii, iii, iv, vii

[MSOI13] MAKINO Y., SUGIURA Y., OGATA M., INAMI M.: Tangential force sensing system on forearm. In *Proceedings of the 4th Augmented Human International Conference* (New York, NY, USA, 2013), AH '13, ACM, pp. 29–34. ii

[PVGO96] PROESMANS M., VAN GOOL L., OOSTERLINCK A.: One-shot active 3d shape acquisition. In *Proceedings of the International Conference on Pattern Recognition (ICPR '96) Volume III-Volume 7276 - Volume 7276* (Washington, DC, USA, 1996), ICPR '96, IEEE Computer Society, pp. 336–. ii

[RHKK11] RHEE T., HWANG Y., KIM J. D., KIM C.: Real-time facial animation from live video tracking. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2011), SCA '11, ACM, pp. 215–224. i, ii

[SFP99] SCHEIRER J., FERNANDEZ R., PICARD R. W.: Expression glasses: A wearable device for facial expression recognition. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 1999), CHI EA '99, ACM, pp. 262–263. i, ii

[SPIF07] SALZMANN M., PILET J., ILIC S., FUA P.: Surface deformation models for nonrigid 3d shape recovery. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 8 (Aug. 2007), 1481–1487. iii

[VWB*12] VALGAERTS L., WU C., BRUHN A., SEIDEL H.-P., THEOBALT C.: Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 187:1–187:11. ii

[WBGB16] WU C., BRADLEY D., GROSS M., BEELER T.: An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph. 35*, 4 (July 2016), 115:1–115:12. i, ii

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 Papers* (New York, NY, USA, 2011), SIGGRAPH '11, ACM, pp. 77:1–77:10. i, ii

[Wil90] WILLIAMS L.: Performance-driven facial animation. In *Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1990), SIGGRAPH '90, ACM, pp. 235–242. i, ii

[WLVGP09] WEISE T., LI H., VAN GOOL L., PAULY M.: Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2009), SCA '09, ACM, pp. 7–16. ii

[ZSCS04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S. M.: Spacetime faces: High resolution capture for modeling and animation. *ACM Trans. Graph. 23*, 3 (Aug. 2004), 548–558. ii