Improvement of Lip Extraction Method Using Interpolation Method

Etsuro Nakamura\textsuperscript{a}, Tsuyoshi Takahashi\textsuperscript{b}, Yoichi Kageyama\textsuperscript{a,}\textsuperscript{*}, Masaki Ishii\textsuperscript{b}, Makoto Nishida\textsuperscript{c}, and Motonari Shirasu\textsuperscript{d}

\textsuperscript{a}Akita University, 1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan
\textsuperscript{b}Akita Prefectural University, 84-4 Tsuchiya-Ebinokuchi, Yurihonjo, Akita 015-0055, Japan
\textsuperscript{c} The Open University of Japan, 1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan
\textsuperscript{d} Japan Business Systems, Inc., 1-23-1 Toranomon, Minato-ku, Tokyo 105-6316, Japan

*Corresponding Author: kageyama@ie.akita-u.ac.jp

Abstract

Feelings of amusement improve the quality of life (QOL). However, in the authors’ previous studies, it was revealed that the variance value of lip movement increases when people feel amusement. Therefore, it is important to extract the lip shape accurately and automatically for quantifying amusement feelings and the QOL in daily life. A lip shape extraction method using a feedforward neural network and correction method using a shadow-removing method for the lip shape extraction method (previous method) was proposed. The previous method can accurately extract the lip shape and solve the problem that the lip is extracted excessively from the shadow around the lip. However, the previous method has the problem that the entire lip area is not extracted because of the shadow or highlights. In this work, improvement of the lip extraction method using an interpolation method is proposed. In the experiments conducted, the proposed method was less susceptible to lighting and provided more-satisfactory lip area extraction than the previous method.

Keywords: lip, human interface, feedforward neural network.

1. Introduction

In recent years, such problems as increasing costs of nursing care and medical expenses in the super-aging society in Japan have intensified. Maintaining and improving people’s quality of life (QOL) can solve these problems in the super-aging society in Japan. Therefore, people have started taking into consideration the QOL to solve the problems\textsuperscript{(1)}. However, it is known that feeling amused improves the QOL. In previous studies, it was revealed that the variance value of lip movement increases when people feel amusement\textsuperscript{(2)}. The psychological change is delicate. Therefore, it is important to extract the lip shape accurately for quantifying the amusement feeling and QOL.

As conventional methods, a face feature points extraction method using Convolutional Neural Network (CNN)\textsuperscript{(3)}, and face detection method using Recurrent Neural Network (RNN) and CNN\textsuperscript{(4)}. These methods can extract the lip feature points from face image. Furthermore, a face parts extracted method based on machine learning using face shape vector, face image, and trained function\textsuperscript{(5)} were proposed by other research groups. However, many training data are needed for machine learning. In addition, the lip extraction method based on the face shape vector cannot extract the lip shape when part of the face is hidden.

A lip shape extraction method using a feedforward neural network (FFNN)\textsuperscript{(6)} and a correction method using a shadow-removing method for the lip shape extraction method\textsuperscript{(7)} (previous method) was proposed. The previous method can accurately extract the lip shape using a few amount training data, and can extract lip shape even if part of the face is hidden. Furthermore, the previous method solves the problem that the lip is extracted excessively from the shadow around the lip. However, in the previous method, the entire lip area is not extracted. The reason is that the features of the lip and skin become similar because of the shadow or highlights.

In this work, improvement of the lip extraction method using an interpolation method (proposed method) is proposed. The proposed method can extract lip area is not extracted using the previous method. The proposed method, the previous method and the lip extraction method based on feature points were compared to clarify the usefulness of the proposed method.
2. Data Acquisition

Image data of seven subjects’ (A–G, all of Asian descent) pronunciation were taken using a web camera (Logicock C922 PRO STREAM WEBCAM) from the front in three kinds of lighting environment (I–III). The data acquisition environment is shown in Fig. 1, and the three kinds of lighting environment are shown in Table 1. A pronouncing command “o/fu/ro” (meaning “bath” in English) was used. The given command was used for two reasons.

- The word is familiar.
- Shadows tend to appear under the lower lip.

The data used in this study were acquired in accordance with the ethical regulations concerning studies involving humans at Akita University, Japan.

3. Proposed Method

3.1 Outline of the Proposed Method

As shown in Fig. 2, the proposed method consists of three processes. The first process is the lip shape extraction method using an FFNN. The second process is the correction method using a shadow-removing method. These two processes used the previous method for lip shape extraction. The third process is the interpolation process for the extracted lip shape. It can interpolate the unextractable lip area using the previous method.

3.2 Lip Extraction Process

Lips were extracted using a lip extraction method(6) based on an FFNN. First, the first frame was obtained from all frames for training the data. Second, the rima oris was extracted from the first frame. Third, features were extracted, six patterns each, from the lip and skin based on the rima oris for training the FFNN. Finally, the FFNN was trained, and the lip shape was extracted from all frames of one utterance using the trained FFNN.

3.3 Shadow-Removing Process

The shadow-removing process(7) was executed based on the L* (lightness value) of an L*a*b* color space(8), because the shadow area had low lightness. First, the threshold was calculated based on the average L* on the lip area. Next, the shadow areas were extracted from the lip area using the threshold. Finally, the shadow area was removed from the extracted lip area.

3.4 Interpolation Process for Extracted Lip Shape

The interpolation process consists of four processes (a)–(d).

(a) Contour pixel extraction process

Contour pixels of the lip were extracted from some lip pixels in the lip extraction image, based on the following criteria of lip contour, Class 1–4.

- Class 1: The lip pixels have skin areas in all pixels of the left and up directions.
- Class 2: The lip pixels have skin areas in all pixels of the left and bottom directions.
Class 3: The lip pixels have skin areas in all pixels of the right and up directions.

Class 4: The lip pixels have skin areas in all pixels of the right and bottom directions.

Fig. 3 shows an example of the result for the contour pixels extracted.

(b) Moving process for contour pixels

Lip shape was interpolated by linking lip contour pixels. However, contours of the lip extracted correctly were interpolated to the wrong shape. Therefore, the contour pixels of the lip extracted were moved on the inside of the lip contours. Lip contour pixels in each class were moved using the following process, Class 1–4.

- Class 1: All lip contour pixels were moved by (1, 1) pixels.
- Class 2: All lip contour pixels were moved by (1, -1) pixels.
- Class 3: All lip contour pixels were moved by (-1, 1) pixels.
- Class 4: All lip contour pixels were moved by (-1, -1) pixels.

Fig. 4 shows an example of the contour pixel moving result.

(c) Linking process for contour pixels

Lip shape was interpolated by linking lip contour pixels classified in the same class. Fig. 5 shows an example of the linking process for contour pixels. First, x coordinate values of all contour points were calculated. Second, the differences (dis) between the maximum and minimum values of x coordinate values were calculated in each class. Third, the nth pixel (Pn) and pixels in the same class as Pn and within a distance of half dis pixels from Pn were linked. Finally, the above processing was performed for all classes.

(d) Interpolation process of lip shape

First, a binary image in which the lip area and connecting line were white pixels and other areas were black pixels was made. Next, black pixels between white pixels were reset to white pixels. Fig. 6 shows an example of the image before and after the interpolation process adaptation.

4. Evaluation Experiment

4.1 Mask Image and Evaluation Index

Mask images of the seven subjects' lip areas were created to compare the proposed method and previous method. Mask images were created by one operator based on the following lip criteria i)–iii).

i) The area that exists in the mouth part of the face and is reddish.

ii) The area of the mouth corner.

iii) The area where the upper lip and the lower lip are in contact and rima oris.

\( F\text{-score} \) was used to compare of the proposed method with the previous method. First, precision and recall were calculated as per (1) and (2). Next, the \( F\text{-score} \) was calculated as per (3) using precision and recall values.

\[ TP = \text{number of pixels with which the lip is correctly classified as the lip}, \quad FP = \text{number of pixels with} \]

\( \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
which the skin is incorrectly classified as the lip, and $FN$ is number of pixels with which the lip is incorrectly classified as the skin.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{2}
\]

\[
F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \tag{3}
\]

In addition, the proposed method was compared to the lip extraction method based on the face shape vector is implemented using the machine learning library dlib\(^{10}\). Officially distributed face detector\(^{11-13}\) was used in the dlib. The dlib can get 20 lip feature points from face image. Intersection-over-Union (IoU)\(^{14}\) was used to compare of the proposed method with the dlib. IoU was calculated as per (4), and two rectangles was used for calculation of IoU. One is the rectangle that circumscribes the lip area of the mask image, while the other is one that circumscribes the lip area extracted automatically. “Area of overlap” is the logical AND of the two rectangles, and “Area of union” is the logical OR of the two rectangles.

\[
\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{4}
\]

### 4.2 Calculation of Evaluation Index

(a) Calculation of F-score
First, lip areas were extracted for seven subjects’ image data using the proposed method and previous method. Second, the $F\text{-score}$ was calculated using the extracted lip area and the mask image. Third, the average value of the $F\text{-score}$ ($F_{\text{ave}}$) was calculated for each lighting environment of each subject. Finally, the average $F_{\text{ave}}$ and standard deviation $F_{\text{ave}}$ were calculated for each subject, and the results of the proposed method and previous method were compared.

(b) Calculation of IoU
First, lip areas and lip feature points were extracted using the proposed method and dlib from seven subjects’ image data, respectively. Second, the $\text{IoU}$ was calculated using the mask image for lip area and lip feature points, respectively. Third, the average value of the $\text{IoU}$ ($\text{IoU}_{\text{ave}}$) was calculated for each lighting environment of each subject. Finally, the average $\text{IoU}_{\text{ave}}$ and standard deviation $\text{IoU}_{\text{ave}}$ were calculated for each subject, and the results of the proposed method and dlib were compared.

### 5. Experimental Result

#### 5.1 Evaluation Results of Average $F_{\text{ave}}$
Table 2 shows the calculation result of the average $F_{\text{ave}}$ for each method. In all subjects, the average $F_{\text{ave}}$ of the proposed method was higher than that of the previous method, and the best value was 0.9477. The proposed method can improve the average $F_{\text{ave}}$ by 0.0436 on average. Fig. 7 shows an example of the resulting image. The results suggest that the proposed method is more useful for lip extraction than the previous method.

#### 5.2 Evaluation Results of Standard Deviation $F_{\text{ave}}$
Table 3 shows the calculation result of the standard deviation $F_{\text{ave}}$ for each method. In all subjects, the standard deviation $F_{\text{ave}}$ of the proposed method was lower than that of the previous method. The results of the average $F_{\text{ave}}$ becoming higher and the standard deviation $F_{\text{ave}}$ becoming
lower suggest that \( F_{\text{ave}} \) was improved in each lighting environment. The results suggest that the proposed method is useful for lip extraction and reduces the effects of lighting compared with the previous method.

### 5.3 Evaluation Results of Average IoU\text{ave}

Table 4 shows the calculation result of the average IoU\text{ave} for each method. In 5 out of 7 subjects, the average IoU\text{ave} of the proposed method was higher than that of the dlib, and the best value was 0.9147. Furthermore, subject A’s difference of the average IoU\text{ave} was small. The results suggest that the proposed method is useful for lip extraction. On the other hand, subject F’s average IoU\text{ave} of the propose method is lower than the dlib because the lip shape extracted using an FFNN\textsuperscript{(6)} was over corrected by the shadow removing method\textsuperscript{(7)} in the proposed method. Therefore, by improving the shadow removing method, lip shape extraction accuracy is improved.

### 5.4 Evaluation Results of Standard Deviation IoU\text{ave}

Table 5 shows the calculation result of the standard deviation IoU\text{ave} for each method. In 6 out of 7 subjects, the standard deviation IoU\text{ave} of the proposed method was lower than that of the dlib. The results of the average IoU\text{ave} becoming higher and the standard deviation IoU\text{ave} becoming lower suggest that proposed method can extract the lip shape with accuracy equal to or better than the dlib.

### 6. Conclusions

An interpolation method was proposed for the lip extraction method using an FFNN, and the accuracy of lip extraction was improved. The experiments demonstrated that i) the interpolation method could solve the problem of the previous method that the entire lip area is not extracted, ii) the proposed method is less susceptible to lighting than the previous method and can extract lip area more satisfactorily than the previous method, and iii) the proposed method can extract the lip shape with accuracy equal to or better than the dlib.

### Acknowledgment

The authors would like to thank Japan Business Systems, Inc. for its assistance in conducting this study. This study has been supported by JSPS KAKENHI Grant Number JP19K12909.

### References

(6) Etsuro Nakamura, Tsuyoshi Takahashi, Yoichi Kageyama, Masaki Ishii, and Makoto Nishida :


(8) Mikio Takagi and Haruhisa Shimoda: “Handbook of Image Analysis”, Tokyo, Japan, University of Tokyo Press, 2004

(9) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, (Yusuke Iwazawa et al., Trans.): “DEEP LEARNING”, Tokyo, Japan, Dowango, Inc., 2018


(14) Kazuya Hujita and Ayumu Takahara: “Implemented deep learning”, Tokyo, Japan, Ohmsha, 2016