Robust AAM-Based Audio-Visual Speech Recognition against Face Direction Changes

Yuto Komai
komai@me.cs.scitec.kobe-u.ac.jp

Nan Yang
yangnan@me.cs.scitec.kobe-u.ac.jp

Tetsuya Takiguchi
takigu@kobe-u.ac.jp

Yasuo Ariki
ariki@kobe-u.ac.jp

Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan

ABSTRACT

As one of the techniques for robust speech recognition under noisy environments, audio-visual speech recognition (AVSR) using lip dynamic scene information together with audio information is attracting attention, and the research has advanced in recent years. However, in visual speech recognition (VSR), when a face turns sideways, the shape of the lip as viewed from the camera changes and the recognition accuracy degrades significantly. Therefore, many of the conventional VSR methods are limited to situations in which the face is viewed from the front. This paper proposes a VSR method to convert faces viewed from various directions into faces that are viewed from the front using Active Appearance Models (AAM). In the experiment, even when the face direction changes about 30 degrees relative to a frontal view, the recognition accuracy improved significantly.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing—Speech recognition; I.5.4 [Pattern Recognition]: Applications—Computer vision, Signal processing

General Terms

Algorithms, Experimentation

Keywords

audio-visual, speech recognition, face direction

1. INTRODUCTION

In recent years, audio speech recognition (ASR) software for PCs and mobile phones has become widely used and attracts attention as a hands-free technology replacing the input from a keyboard. However, in current ASR technologies, the recognition performance degrades under noisy environments, which is a significant problem in regard to making practical use of it in speech recognition.

Human beings use a variety of information comprehensively when understanding the content of an utterance. For example, when it is hard to hear the voice, the listener pays attention to the speaker's lip movement and tries to understand what is being said. Conversely, in the case where the lip movement does not match with the speech, he may misunderstand what is being said. This is called the McGurk effect, and it indicates that phonological perception is not decided only by audio information but also by visual information, such as lip movement. Thus, it is important for speech recognition to integrate lip information and audio information.

A technology to recognize speech content from lip motion is called visual speech recognition (VSR). VSR is not influenced by noise, whereas ASR is sensitive to noise and its recognition rate degrades significantly under noisy environments. Therefore, as one of the techniques for robust speech recognition under noisy environments, audio-visual speech recognition (AVSR), using VSR together with ASR, is attracting attention [8].

However, in VSR, when a face turns sideways, the shape of the lip, viewed from a camera fixed in front of the user, changes, and the recognition accuracy degrades significantly. Thus, many of the conventional VSR approaches are limited to situations in which the face is viewed from the front. Therefore, there is a great need to be able to recognize visual speech from arbitrary face directions.

VSR locates the lip ROI (Region of Interest) and extracts the lip features. For detection of lip ROI, traditional image processing techniques, such as color segmentation [6] and edge detection [3], were employed, along with statistical modeling techniques, such as Snakes [9], Active Shape Models (ASM) [10] and Active Appearance Models (AAM) [1]. For the visual features, appearance-based features, such as PCA [11] and DCT [4], and shape-based features, such as the width and height of the lip [12], were employed. Furthermore, a combination of both appearance and shape features, such as AAM parameters [5] have been employed recently.

In regard to research of VSR from various face directions, there is a method that trains the transformation matrices from the profile view to the frontal view and transforms the faces from side to front [7]. However, this technique re-

MM'12, October 29–November 2, 2012, Nara, Japan.

quires transformation matrices in each direction. Thus, it is difficult to recognize visual speech with arbitrary face directions. In this paper, we propose a method to extract the lip area automatically in various face directions and to recognize visual speech by converting the sideways lip figure into a frontal one using Active Appearance Models (AAM). The experiment results show that the proposed method provides better performance in comparison with the conventional approaches.

2. OVERVIEW OF VISUAL SPEECH RECOGNITION

First, the face area is detected based on AdaBoost, using the Haar-like features on the input image. This is because the extraction of the feature points using AAM greatly depends on the initial search area. Therefore, the extraction accuracy of the feature points is improved by applying the detected face area to AAM as an initial search area. After detecting the face area, AAM is applied to the detected face area, and the facial feature points are extracted. Then AAM generates the model parameters most similar to the input image. The speaker's face direction is estimated from the generated parameters using the method described in Section 4.2. After estimating the face direction, using the method described in Section 4.3, a face in any direction is converted to a frontal face (we call this operation "normalization"). Finally, the lip features are extracted, and the visual speech is recognized using HMMs.

The lip feature employed is an AAM model parameter [5] that includes shape information and texture information. In this paper, AAM is applied to the whole face area in order to estimate the face directions accurately, but the AAM model parameters also contains information other than the lip and its movement when whole face AAM is applied. Therefore, after normalization of face direction, some dimensions that include the lip information predominantly in the AAM parameters are extracted and recognized. These dimensions are extracted, from among all the dimension combinations, as the best combinations with the highest recognition accuracy of the visual speech.

In this paper, the audio signal is converted to MFCCs (mel-frequency cepstral coefficients) that are commonly used in a standard speech recognition system. In training, audio and visual HMMs are independently constructed using each feature vectors extracted from the same movie. In testing, a final likelihood is calculated using the late integration of likelihoods from audio HMMs and visual HMMs as follows:

$$L_{A+V} = (1-\alpha)L_A + \alpha L_V, \quad 0 \leq \alpha \leq 1 \tag{1}$$

where L_A and L_V are likelihoods of audio and visual features, respectively. α is the combination weight.

3. ACTIVE APPEARANCE MODELS

AAM is a technique used to express a facial model using low-dimensional parameters. The subspace is constructed by applying PCA to shape and texture of face feature points.

The shape vector $\mathbf{s}$, the feature points on the face images, and mean shape $\bar{\mathbf{s}}$ are computed from the training image set. The inner texture of $\mathbf{s}$ is normalized to mean shape. The shape vector $\mathbf{s}$ and the texture vector $\mathbf{g}$ are given: $\mathbf{s} = (x_1, y_1, ..., x_n, y_n)^T$, $\mathbf{g} = (g_1, ..., g_m)^T$, where x_i, y_i $(i \leq n)$ are the coordinates of the feature points. g_j $(j \leq m)$ is the intensity value at each pixel in $\bar{\mathbf{s}}$, and mean intensity value $\bar{\mathbf{g}}$ can be computed from the training image set. $\mathbf{s}$ and $\mathbf{g}$ are expressed by using eigenvector matrices $\mathbf{P_s}$ and $\mathbf{P_g}$, obtained by applying PCA to deflection from $\bar{\mathbf{s}}$ and $\bar{\mathbf{g}}$, as shown in Eq. (2) and Eq. (3).

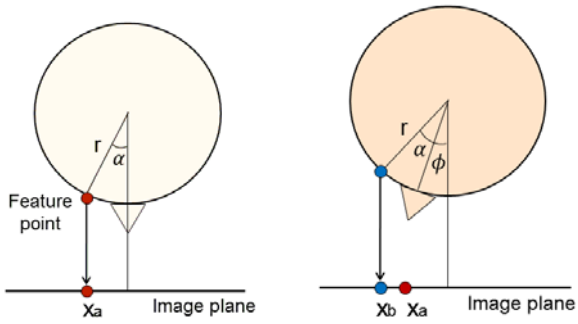
$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s} \tag{2}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \tag{3}$$

$\mathbf{b_s}$ and $\mathbf{b_g}$ are called the shape parameter and the texture parameter, respectively, and shape vector $\mathbf{s}$ and texture vector $\mathbf{g}$ are converted to each of them, respectively. Moreover, $\mathbf{b_s}$ and $\mathbf{b_g}$ are combined and reduced as shown in Eq. (4) by applying PCA because there is a correlation in shape and texture.

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s}^T(\mathbf{s} - \bar{\mathbf{s}}) \\ \mathbf{P_g}^T(\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} = \mathbf{Qc} \tag{4}$$

where $\mathbf{W_s}$ is the matrix that normalizes the difference of the unit of the shape vector and the texture vector. $\mathbf{Q}$ is an eigenvector matrix. $\mathbf{c}$ is a vector of combined shape and texture parameters. This parameter controls both shape and texture. Thus, it becomes possible to treat shape and texture together by controlling only parameter vector $\mathbf{c}$.



(a) Face direction = 0 (b) Face direction = ϕ

Figure 1: Schematic of a face viewed from the head top

4. NORMALIZATION OF THE FACE DIRECTION

A normalization method of the face direction was introduced in [2], and the extended approach is proposed in this paper, where a multiple regression model is used to estimate the visual feature instead of a single regression. Each regression model in our method depends on a phoneme class.

4.1 Regression model

Fig. 1 shows a schematic of a face viewed from the top of the head. Face is regarded as a sphere with radius r. A vertical line is drawn to the image plane from the center of the head. Then, the facial feature point at the angle α from the vertical line is projected onto the coordinates Xa of the image plane as shown in Fig. 1 (a). Furthermore, the facial feature point is projected onto the image plane Xb when the face rotates by the angle ϕ as shown in Fig. 1 (b). Δx, the distance between two feature coordinate points, is expressed as shown in Eq. (5).

$$\begin{aligned} \Delta x &= x_b - x_a = r\sin(\phi+\alpha) - r\sin\alpha \\ &= r\sin\phi\cos\alpha + r\cos\phi\sin\alpha - r\sin\alpha \end{aligned} \tag{5}$$

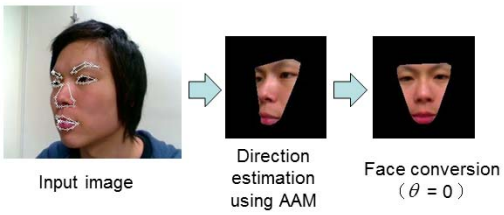


Figure 2: Example of conversion from directional face to frontal face

A regression model can be derived considering r and α as constants, as shown in Eq. (6).

$$\mathbf{c} = \mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi \tag{6}$$

where, $\mathbf{c_0}$, $\mathbf{c_1}$, and $\mathbf{c_2}$ are the regression coefficient vectors estimated from the training data.

4.2 Estimation of the face direction

When AAM is applied to a new input image with no information of face direction, parameter $\mathbf{c}'$ is generated. Then, the direction ϕ can be estimated as shown in Eq. (7) using Eq. (6).

$$\begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} = \mathbf{B}^{+}(\mathbf{c}^{'} - \mathbf{c_0}) \tag{7}$$

where $\mathbf{B}^{+}$ is the pseudo inverse matrix of $(\mathbf{c_1}\ \ \mathbf{c_2})$. Therefore, the direction ϕ is estimated as shown in Eq. (8) using $\cos\phi$ and $\sin\phi$ in Eq. (7).

$$\phi = \tan^{-1}(\sin\phi/\cos\phi) \tag{8}$$

4.3 Converting of the directional face to frontal face

When AAM is applied to the input image, parameter $\mathbf{c}'$ is generated using AAM and face direction ϕ is obtained using Eq. (8). Then, the residual vector $\mathbf{c_{res}}$ is estimated as shown in Eq. (9).

$$\mathbf{c_{res}} = \mathbf{c}^{'} - (\mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi) \tag{9}$$

The directional face is expressed as shown in Eq. (10) using Eq. (9).

$$\mathbf{c_{new}} = \mathbf{c_0} + \mathbf{c_1}\cos\theta + \mathbf{c_2}\sin\theta + \mathbf{c_{res}} \tag{10}$$

If $\theta = 0$, a face direction is converted to the front. Fig. 2 shows the result of the conversion from directional face to frontal face.

4.4 Multiple regression model

In this paper, a multiple regression model of Eq. (6) is estimated in order to decrease the variation mismatch between frontal face and directional face. The i-th regression model is represented as follows:

$$\mathbf{c^i} = \mathbf{c_0^i} + \mathbf{c_1^i}\cos\phi + \mathbf{c_2^i}\sin\phi \tag{11}$$

Each regression model is estimated using the only training data for a phoneme. In this paper, six regression models are estimated using the training data for the Japanese vowel: /a/, /i/, /u/, /e/, /o/ and the nasal /N/, respectively.

In the process of conversion to frontal face, first, $\mathbf{c_{in}}$ is obtained by applying AAM to the test image. Next, the face direction θ is estimated using $\mathbf{c_{in}}$ according to Eq. (8). The optimal regression model is selected so that the minimum distance between $\mathbf{c^i}(\theta)$ and $\mathbf{c_{in}}$ is achieved as follows.

$$\hat{\mathbf{i}} = \underset{\mathbf{i}}{\operatorname{argmin}} ||\mathbf{c^i}(\theta) - \mathbf{c_{in}}|| \tag{12}$$

Then, as described in Section 4.3, the face direction is converted to the front.

Table 1: Visual recognition rates [%] without normalization of face direction

	front	15 degrees	30 degrees
c parameter	80.67	13.39	1.30

Table 2: Visual recognition rates [%] with normalization of face direction

	front	15 degrees	30 degrees
c parameter (single regression)	78.67	54.32	42.35
c parameter (multiple regression)	79.56	54.72	49.37
DCT	72.77	50.49	47.48

5. EXPERIMENT

5.1 Experimental condition

Two subjects spoke ATR phoneme-balanced words (216 words)×10 sets with the frontal face, the same 216 words ×1 set with 15-degree face and 30-degree face, respectively. Resolution was 320×240 pixels, and the frame rate was 30 fps.

The leave-one-out method was applied to 216 words×10 sets, where 216 words×9 sets with the frontal face were used for training HMMs, the remaining one set with the frontal face and the 216 words with the directional faces were used for test, and the recognition rate was the average over the 10 sets. Monophone HMMs were constructed with 5 states and 16 mixtures.

The number of AAM training images was 108, and the number of feature points on each image was 63. As a result of feature extraction described in Section 3, the AAM parameter for two subjects was reduced to 5 dimensions and 9 dimensions, respectively, for 95 % of the cumulative proportion. Including the AAM parameter, its delta and delta-delta parameters, were finally used as the visual features. 12-dimensional MFCC parameters, along with their delta and delta-delta parameters, were used as the audio features.

5.2 Experimental results

Table 1 shows the recognition rates for only visual features without normalization of face direction. "Front" indicates the recognition rate of the frontal face. "15 degrees" and "30 degrees" indicate the recognition rate of the 15-degree face and the 30-degree face, respectively. As shown in Table 1, although a high recognition rate is obtained in "front", the recognition rates except for "front" degrade seriously. This is because the shape of the lip viewed from the camera changes in directional face. Therefore, the recognition rates are affected seriously.

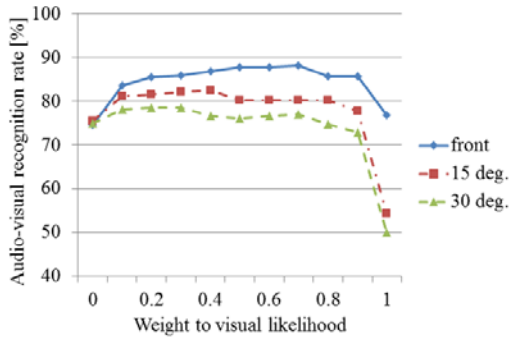


Figure 3: Audio-visual recognition results at SNR of 20 dB

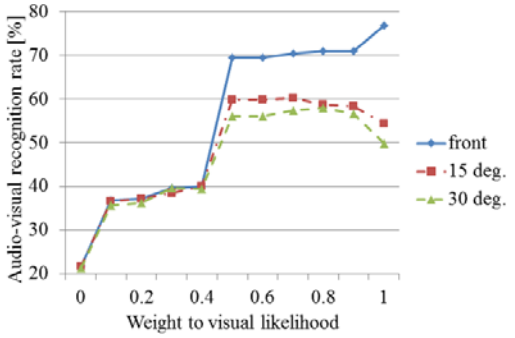


Figure 4: Audio-visual recognition results at SNR of 0 dB

Table 2 shows the recognition rates for only visual features with normalization of face direction. The recognition rate of "15 degrees" improved by about 41.3 points and the recognition rate of "30 degrees" improved by about 48.1 points compared with Table 1. Thus, it was confirmed that the proposed method is effective for directional face, and the performance of the multiple regression is better than that of the single regression. However, the recognition rate of "30 degrees" is low by about 5.35 points compared with "15 degrees". One of the reasons is that the extraction accuracy of the feature points using AAM degrades when the face direction becomes large. Moreover, when converting directional faces, there is a possibility that the lip information is collapsed slightly and the recognition rate degrades.

Table 2 also shows the comparison with the performance of the AAM-based features with that of the conventional 2D DCT-based features. 15 low-frequency components of the 2D DCT feature were selected, and including the DCT parameter, its delta and delta-delta parameters, were finally used as the visual features. As shown in Table 2, it was confirmed that the AAM-based features are more effective than the conventional DCT-based features.

In order to integrate the visual result with the audio result under noisy environments, the likelihoods from visual HMMs and audio HMMs were integrated according to Eq. (1). Fig. 3 and Fig. 4 show the audio-visual recognition results at SNRs of 20 dB and 0 dB, respectively. The combination weight was increased by 0.1 from 0.0 to 1.0, where the weight 0 corresponds to the audio feature only, and 1 to the visual feature only. As shown in both figures, the recognition rate is improved by taking the optimum value of the weight. Although the recognition rate using only audio HMMs greatly decreased in the strong noisy environment at SNR of 0 dB, it could be improved by increasing the weight to the image.

6. CONCLUSION

We proposed the method to recognize visual speech with face directions by converting directional faces into the frontal faces. The experimental results showed that the recognition rate of the directional face was improved in comparison with that without converting the face direction. Also, it could be confirmed that the recognition rate is improved in comparison with that for the only audio feature by integrating the visual feature and audio feature under noisy environments. Future work will include the recognition of utterances spoken by more people, expansion to continuous speech recognition, recognition of speech with spontaneous tone.

7. ACKNOWLEDGMENTS

This research was supported in part by MIC SCOPE.

8. REFERENCES

[1] T. Cootes. Active appearance model. In Proc. European Conference on Computer Vision, number 2, pages 484–498, 1998.

[2] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. Image and Vision Computing, 20(9-10):657–664, August 2002.

[3] Y.-P. Guan. Automatic extraction of lip based on wavelet edge detection. In Proc. SYNASC, pages 125–132, 2006.

[4] H. Jun and Z. Hua. Research on visual speech feature extraction. In Proc. ICCET, pages 499–502, 2009.

[5] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden. Improving visual features for lip-reading. In Proc. AVSP, pages 142–147, 2010.

[6] W. Lirong, X. Jing, and Z. Yanyan. Research of visual features detection and tracking methods about audio-visual bimodal speech recognition. In Proc. IFITA, pages 204–207, 2010.

[7] P. Lucey, G. Potamianos, and S. Sridharan. An extended pose-invariant lipreading system. In Proc. AVSP, 2007.

[8] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, 2004.

[9] M. Ramage and E. Lindsay. Wrapping snakes for improved lip segmentation. In Proc. ICASSP, pages 1205–1208, 2009.

[10] K. Sum, W. Lau, S. Leung, A. WC, Liew, and K. W. Tse. A new optimization procedure for extracting the point-based lip contour using active shape model. In Proc. ICASSP, pages 1485–1488, 2001.

[11] M. Tomlinson, M. Russell, and N. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In Proc. ICASSP, pages 821–824, 1996.

[12] T. Yoshida and K. Nakadai. Audio-visual speech recognition system for a robot. In Proc. AVSP, pages 8–13, 2010.