

Mouth Region Localization Method Based on Gaussian Mixture Model

Kenichi Kumatani and Rainer Stiefelhagen

Universitaet Karlsruhe (TH), Interactive Systems Labs, Am Fasanengarten 5,
76131 Karlsruhe, Germany
k_kumatani@ieee.org, stiefel@ira.uka.de

Abstract. This paper presents a new mouth region localization method which uses the Gaussian mixture model (GMM) of feature vectors extracted from mouth region images. The discrete cosine transformation (DCT) and principle component analysis (PCA) based feature vectors are evaluated in mouth localization experiments. The new method is suitable for audio-visual speech recognition. This paper also introduces a new database which is available for audio visual processing. The experimental results show that the proposed system has high accuracy for mouth region localization (more than 95 %) even if the tracking results of preceding frames are unavailable.

1 Introduction

Facial feature localization methods have recently undergone much attention. In particular, a mouth feature plays an important role for many applications such as automatic face recognition, facial expression analysis and audio visual automatic speech recognition.

However, automatic mouth localization is especially difficult because of the various changes of its shape and person dependent appearance. In addition, the better localization accuracy and faster response speed should be achieved at the same time. Many systems use skin color, the vertical and horizontal integration of pixel values in a face image[1]-[4]. However those systems are generally not robust for the significant change of illumination conditions.

Some researchers have tried to find the precise lip contour [5]-[7]. However, most of applications don't need it. For example, in audio visual speech recognition, the image of a mouth region is preferred [12].

Lienhart et al. applied the detector tree boosted classifiers to lip tracking [8]. And they showed that their tracking system achieved high accuracy and small execution time per frame. However, we found that their method often fails to localize a mouth area at a frame level. In addition, an eye image is often misrecognized as a mouth. Actually they refined the trajectory of mouth by post-processing approach. Although some errors at a frame level can be recovered by such a post-processing, better accuracy at each frame is of course preferred.

The method based on Gaussian mixture models (GMM) [9][10] is one of the promising approaches since its performance is very high. And it can easily adjust the accuracy and computation cost by configuring the parameters such as a number of

mixtures. In the GMM based methods, the feature vector representation is a main issue for the improvement of the performance.

In this paper, we present a new mouth region localization method based on GMM, which doesn't need prohibitively heavy calculation. This paper is organized as the followings: Section 2 describes the training algorithm of GMM and section 3 describes the new mouth region localization method. Then section 4 presents the database used in experiments. In section 5, experimental results are depicted and discussed.

2 Training Algorithm

First this section defines a mouth template image to be localized. Then feature vectors used in this paper are explained. This paper evaluates two kinds of feature representation: (1) discrete cosine transformation (DCT) based feature vector [9] and (2) principle component analysis (PCA) based feature vector. After that, we describe how to construct GMM of the feature vectors.

2.1 Mouth Template

In order to construct the template images of a mouth, consistent rules for labeling are required. In our system, the width of a mouth region is defined from a left lip corner to the right one. And the height is defined from nostrils to a chin. Accordingly the mouth templates include non-lip area. Figure 1 shows the samples of the template images. By containing non-lip area such as the nostrils, template images can have robust information to locate a mouth region because nostrils don't change largely. If we use a lip image only, a mouth region more often fails to be localized because of the significant change of lip shape. It is noteworthy that the movements of nostrils might have useful information for audio-visual speech recognition. All mouth images are scaled to the same size in training stage which is the average size over the training images. Thus, the original ratio of the width to the height is not kept the same.

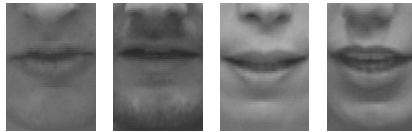


Fig. 1. This figure shows the samples of the mouth template images for training. Note that these samples are scaled to the same size.

2.2 DCT Based Feature Vector

Let I denote the image normalized by histogram equalization and its size is $M \times N$. Then, the 2-D DCT transformation is computed as:

$$D_{u,v} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left[C_i \times C_j \times \cos(u\pi \frac{2m+1}{2M}) \times \cos(v\pi \frac{2n+1}{2N}) \times I_{m,n} \right]. \quad (1)$$

After that, the matrix $D_{u,v}$ is converted into a vector using a zigzag scan.

2.3 PCA Based Feature Vector

Let \mathbf{x} denote the vector converted from the normalized image. This approach calculates the followings.

- (1) The mean of the vectors $\bar{\mathbf{x}}$,
- (2) The covariance matrix \mathbf{C} of the vectors,
- (3) The eigenvectors, Φ_i and the corresponding eigenvalues λ_i of \mathbf{C} (sorted so that $\lambda_i \geq \lambda_{i+1}$),

After these values are calculated, a feature vector is represented as:

$$\mathbf{y} = \Phi^T (\mathbf{x} - \bar{\mathbf{x}}) . \quad (2)$$

where the matrix Φ consists of the t eigenvectors corresponding to the largest eigenvalues.

Although both methods can de-correlate the elements of an image and compress vector size efficiently, the feature vectors obtained by PCA are more dependent of the training data. Therefore, PCA based feature vector can deteriorate if there is a gap between training and test data.

2.4 Training GMM

After feature vectors are calculated, those vectors are classified into k classes by K-mean algorithm. A mixture weight, mean and covariance of a mixture are obtained by dividing the number of samples belonging to the class by the total number of samples and calculating a mean and covariance from the samples in the class, respectively.

2.5 Multi-resolution GMM

To improve the efficiency of the search, we use the multi-resolution framework. In this framework, after the mouth region in a coarse image is located, the estimated location is refined in a series of finer resolution image.

For each training and test image, our system constructs the image pyramid where the higher level has the lower resolution. The base image at level 0 has the original resolution. Subsequent levels are formed by half resolution image sub-sampled from the ones at the one lower level. At each level, GMM is build from the corresponding resolution images. The sizes of feature vectors are kept the same over all levels.

2.6 Mouth Region Localization Based on GMM

Figure 2 shows the basic flow chart of our mouth localization system at each level of the pyramid. Given an input image, the image with the same size as a window is cropped. The window is translated over all pixels of an input image and scaled. Our system scales the window in two ways: (1) with the same original ratio of the width to

the height and (2) without keeping that ratio. The process (2) is important because the mouth template images during the training are resized to the same size without keeping the original ratio due to the change of mouth shape. However, scaling without keeping the ratio leads to extremely heavy computation. Thus, our system changes that ratio within the range from the training data. Then, feature vector is calculated from the cropped image, as mentioned in section 2. Note that the cropped image is normalized by histogram equalization. After that, the likelihood of the feature vector is computed with the GMM. The position and size which gives the maximum likelihood are estimated as the mouth region.

First the above process is performed for the lowest resolution image in the pyramid. In order to avoid converging to the local minima, n candidates are kept at each level. At the next level, the search area is limited based on the n candidates. Those steps are repeated until the candidate is found at the finest resolution.

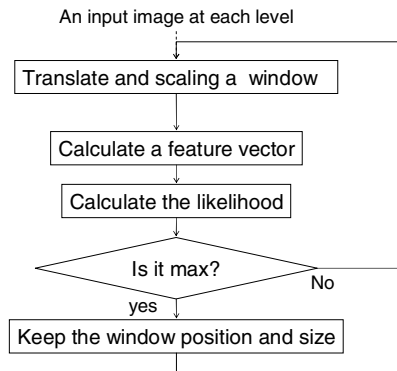


Fig. 2. This figure shows the basic flow chart of the mouth localization system. This process is repeated from the lowest resolution image to the highest resolution image.

3 Database for Experiments

This section describes the specification of the new database we recorded. Though this paper addresses only the video processing, the database contains speech data and is available for audio visual processing.

Figure 3 describes the layout of equipments at the recording. Three pan-tilt-zoom (PTZ) cameras are set at different angles for a subject. A cross talking microphone is put on speaker's ear. Three kinds of video data and two kinds of audio data are recorded. The cameras and microphones are connected to different computers. Audio and video data streams are synchronized with network time protocol (NTP). Figure 4 shows the sample images which are taken at 0, 45 and 90 angles, respectively. The speakers utter English alpha-numeric strings and English sentences extracted from TIMIT database. 39 male and 9 female are recorded.

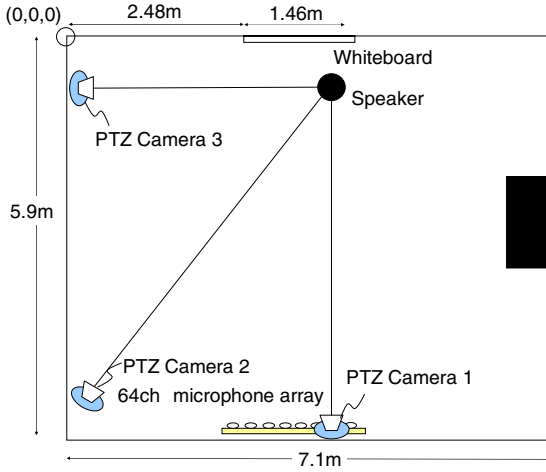


Fig. 3. The layout of equipments at the recording is described

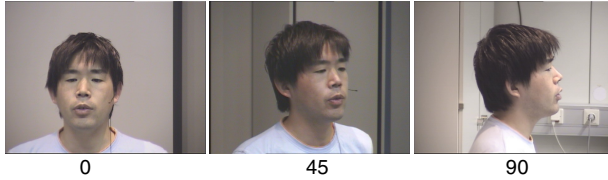


Fig. 4. This figure shows the sample images

4 Experiment

4.1 Experimental Conditions

Table 1 shows the details of experimental conditions. The subjects in test data are not included in training data. In this experiment, images have always only one face. Note that we decide the size of mouth templates at level 0 from the average size over all training images. In this experiment, only frontal faces are used.

4.2 Experimental Results and Discussions

Figure 5 and 6 represent the accuracy of mouth region localization by DCT and PCA based feature vector, respectively. The line 'DN' presents results when the dimension of a feature vector is N . Thus, the line 'D16' indicates results when a 16 dimensional vector is used. The horizontal axis (x-axis) of each figure represents the average distance D_i between the manually labeled points and automatically estimated positions as

$$D_i = \frac{1}{4} \sum_{l=1}^4 \left| \mathbf{p}_{corr}^{(l)} - \mathbf{p}_{est}^{(l)} \right|. \quad (3)$$

where $\mathbf{p}^{(1)} \dots \mathbf{p}^{(4)}$ are positions of an upper left, upper right, bottom left and bottom right of a mouth region, respectively. $\mathbf{p}_{corr}^{(l)}$ and $\mathbf{p}_{est}^{(l)}$ mean the labeled and estimated positions, respectively. The smaller D_i means that system can localize a mouth area more precisely.

Table 1. Experimental conditions

A kind of parameter	Value
Training data	2113 images 30 subjects
Test data	319 images 18 subjects
The number of mixtures	50
The number of candidates kept at each level	8
The size of mouth templates at level 0 (width, height)	65, 105
The maximum pyramid level	3
Dimensions of feature vectors	16, 24, 32, 48, 54
The number of mixtures	50, 80

The vertical axis (y-axis) represents the cumulative probability of x-axis value which is also associated with the accuracy of mouth localization system. For example, figure 5 shows that the mouth regions are correctly estimated with probability 0.97 (accuracy 97 %) by DCT based vector of 54 dimensions when $D_i \geq 15$. On the other hand, Figure 6 shows that the accuracy of 95 % is achieved by PCA based method under the same condition as the above. Comparing Figure 5 with Figure 6, one can see that the mouth can be located more accurately and stably when the PCA based feature is used. However, we found that PCA based method rarely estimate the mouth region far from the correct position. In other words, the completely different area is seldom detected as a mouth region. We consider that those errors occur because the mouth shape is not included in the training data. Note that DCT computation is faster than PCA.

In Figure 7 and Figure8, experimental results are shown when the system uses GMMs with 50 Gaussians (50 mixtures) and 80 Gaussians (80 mixtures). In both figures, the line ‘ $MI(DN)$ ’ stands for GMM with I mixtures and a N dimensional feature vector. Accordingly, the line ‘ $M50(D48)$ ’ indicates the cumulative probabilities when GMM with 50 mixtures and a 48 dimensional feature vector are used. Figure 7 shows experimental results when the DCT based feature vector is used. And results are shown in Figure 8 when the PCA based feature vector is used. Generally by increasing the number of mixtures, we can achieve the better performance for the classification of training data. However, in the case that too many mixtures are used for a few training data, the performance gets worse because of data sparseness. From Figure 8, one can clearly confirm the degradation of the performance when 80 mixtures are

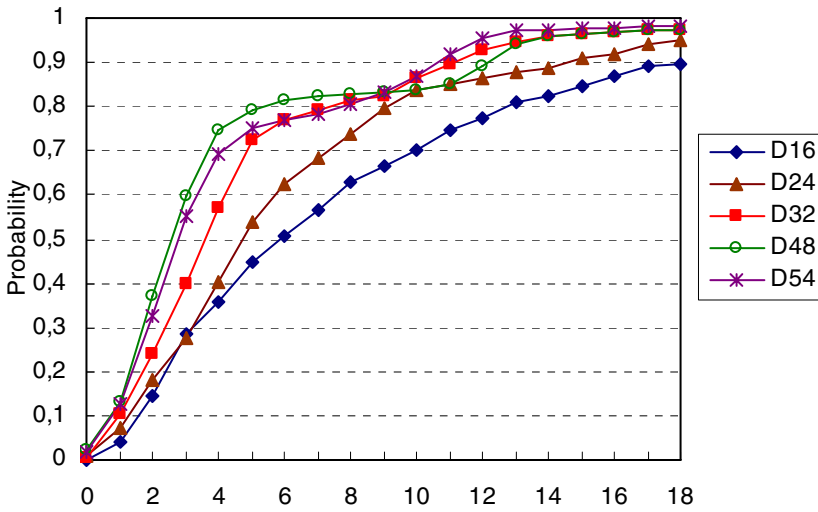


Fig. 5. This figure presents accuracy of mouth localization by DCT based feature vector. In this figure, 'D16' stands for a 16 dimensional feature vector. Accordingly, the line 'D16' (with diamond symbols) indicates the accuracy when a 16 dimensional feature vector is used.

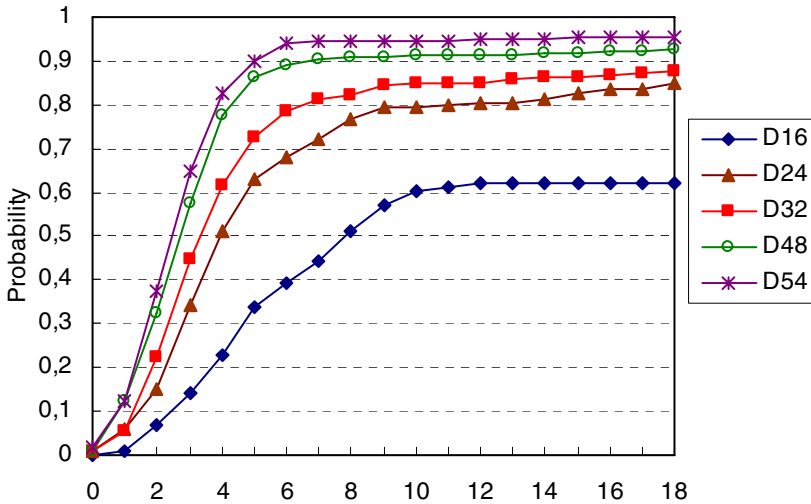


Fig. 6. This figure presents accuracy of mouth localization by PCA based feature vector. In this figure, 'D16' stands for a 16 dimensional feature vector. Accordingly, the line 'D16' (with diamond symbols) indicates the accuracy when a 16 dimensional feature vector is used.

used since it's too much. However, a situation is more complicated when the DCT based feature vector is used. The combination of the number of mixtures and the number of dimension influences the performance. For example, even if the number of mixtures is a little big, the improvement might be obtained by decreasing the number

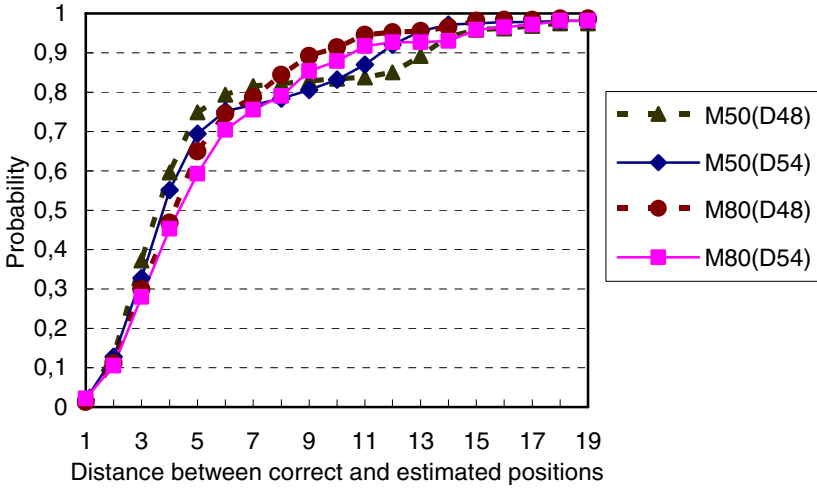


Fig. 7. This figure shows accuracy of mouth localization for the number of mixtures by DCT based feature vector. In this figure, ‘M50(D48)’ stands for GMM with 50 mixtures and a 48 dimensional feature vector. Accordingly, the line ‘M50(D48)’ (the dotted line with triangle symbols) indicates the accuracy when GMM with 50 mixtures and a 48 dimensional feature vector are used.

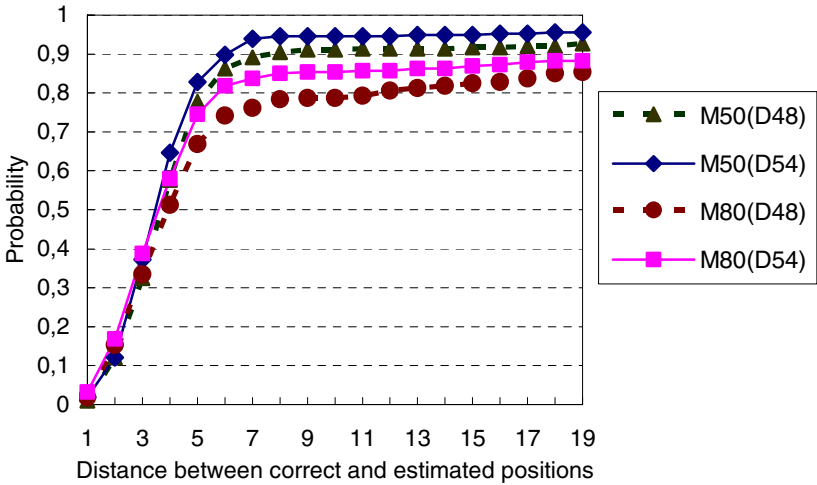


Fig. 8. This figure shows accuracy of mouth localization for the number of mixtures by the PCA based feature vector. In this figure, ‘M50(D54)’ stands for GMM with 50 mixtures and a 54 dimensional feature vector. Accordingly, the line ‘M50(D54)’ (the solid line with diamond symbols) indicates the accuracy when GMM with 50 mixtures and a 54 dimensional feature vector are used.

of dimensions. In fact, when 80 mixtures and 48 dimensions are set (M80(D48)), the best performance is achieved, as shown in Figure 7. Setting too many mixtures also leads to heavy computation.

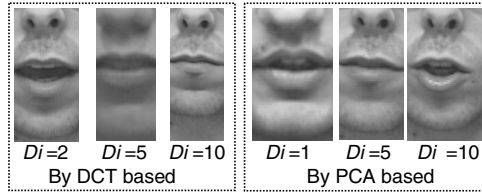


Fig. 9. Examples of result images are depicted. Di is defined in Equation 3 and the same as x -value in Figure 5-8.

Figure 9 shows examples of the mouth images estimated by the DCT and PCA based methods. The DCT based method tends to lose the edge of a chin, as shown in the image above $Di = 10$ of Figure 9, where Di is defined in Equation 3. Note again that Di is also the same as x -value in Figure 5-8. The inaccurate localization of a chin is the main reason why the DCT is less accurate than PCA.

5 Conclusion

We have successfully developed the accurate mouth localization system, which achieved the localization rate 95 % for our database if the average pixel distances Di is more than 6 (see Figure 6). It also proved that PCA based feature can improve the accuracy of the mouth localization. In the future, we are going to embed this method into audio visual speech recognition system.

Acknowledgments

This work was sponsored by the European Union under the integrated project CHIL, Computers in the Human Interaction Loop (<http://chil.server.de>).

References

1. Vladimir Vezhnevets, Stanislav Soldatov, Anna Degtiareva: Automatic Extraction of Frontal Facial Features. Proc. Asian Conf. on Computer Vision, Vol. 2., Jeju (2004) 1020-1025
2. Xingquan Zhu, Jianping Fan, Ahmed K. Elmagarmid: Towards Facial Feature Extraction and Verification for Omni-face Detection in Video/images, Proc. the IEEE Int. Conf. on Image Processing, Vol. 2., New York (2002) 113-116
3. Ying-li Tian, Takeo Kanade, Jeffrey F. Cohn: Lip Tracking by Combining Shape, Color and Motion. Proc. Asian Conference on Computer Vision, Taipei (2000) 1040-1045
4. Selin Baskan, Mehmet Mete Bulut, Volkan Atalay: Projection based Method for Segmentation of Human Face and its Evaluation. Pattern Recognition Letters, Vol. 23., (2002) 1623-1629

5. Haiyuan Wu, Taro Yokoyama, Dadet Pramadihanto, Masahiko Yachida: Face and Facial Feature Extraction from Color Image. Proc. Int. Conf. on Automatic Face and Gesture Recognition, Killington (1996) 345-350
6. Mark Barnard, Eun-Jung Holden, Robyn Owens: Lip Tracking using Pattern Matching Snakes. In Proc. Asian Conf. on Computer Vision, Melbourne (2002) 23-25
7. Juergen Luetttin: Visual Speech and Speaker Recognition. PhD thesis, Department of Computer Science, University of Sheffield (1997)
8. Rainer Lienhart, LuHong Liang, Alexander Kuranov: A Detector Tree of Boosted Classifiers for Real-time Object Detection and Tracking. Proc. IEEE Int. Conf. on Multimedia and Expo, Baltimore (2003) 277-280
9. Jintao Jiang, Gerasimos Potamianos, Harriet J. Nock, Giridharan Iyengar, Chalapathy Neti: Improved Face and Feature Finding for Audio-visual Speech Recognition in Visually Challenging Environments. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 5., Montreal (2004) 873-876
10. Kah-Kay Sung, Tomaso Poggio: Example-based Learning for View-based Face Detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20., (1998) 39-51
11. Gerasimos Potamianos, Chalapathy Neti, Juergen Luetttin, Iain Matthews: Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, 2004.