Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Automatic Visual Feature Extraction for Mandarin Audio-Visual Speech Recognition

Tsang-Long Pao, Wen-Yuan Liao, Tsan-Nung Wu, Ching-Yi Lin
Department of Computer Science and Engineering
Tatung University
Taipei, Taiwan, R.O.C
e-mail: tlpao@ttu.edu.tw, andres@dlit.edu.tw, {g9606005, g9606041}@ms.ttu.edu.tw

***Abstract*—Automatic speech recognition (ASR) by machine has been an attractive research area in past several decades. In recent years, there are many automatic speech-reading systems proposed that utilizing the combination of audio and visual speech features. In this paper, we proposed an automatic visual feature extraction approach to extract the visual features of the lips that can be used in the audio-visual speech recognition system. These features are important to the recognition system, especially in noisy condition. The segmentation of the lip region uses both color and edge information. We then establish a set of visual speech parameters and incorporate them into the recognizer. The WD-KNN classifier is used as the recognition engine in this paper. We present recognition performance using various visual features to explore their impact on the recognition accuracy. These features include the geometric and the motion of the lip. The experimental results based on Mandarin databases demonstrate that the visual information is highly effective for improving the recognition performance.**

***Keywords*—audio-visual speech recognition, audio speech feature, visual speech feature , WD-KNN classifier.**

I. INTRODUCTION

It is well known that people with normal hearing also use the lip-reading as a complementary information source during the conversation, especially when the listening environment is noisy or difficult. So we can say that the human speech perception is a bimodal process involving the analysis of the acoustic and visual signals. It was also proven that the presentation of bimodal speech was considered to be more intelligible than audio-only speech.

Automatic speech recognition (ASR) by computer is an attractive research area in the past several decades. Most of the ASR systems use the acoustic speech signal only and ignore the visual speech cues. Although acoustic-only based ASR systems yield quite excellent results in the laboratory or office environment, the recognition performance in real world is still not good enough. Without the visual features, it will limit the performance of the ASR in the presence of background noise and restrict its usability. Some research had been proposed to increase the robustness of ASR systems under difficult listening conditions. [2-3],[7-8],[10-12]

In recent year, there are some automatic speech-reading systems proposed which use the combination of audio and visual speech features in the recognition process. For such systems, the objective of the audio-visual speech recognizer is to improve the recognition accuracy, particularly in difficult condition. Visual feature extraction and audio-visual fusion are the two most important problems to deal with in this kind of bimodal recognition system. Thus, the audio-visual speech recognition is a work that combined the disciplines of image processing, visual-speech recognition and multi-modal data integration.

To let a computer perform the speech-reading identification, two problems need to be addressed. Firstly, an accurate and robust audio and visual speech feature extraction algorithm needs to be developed. Secondly, fusion method of the two separate information sources needs to be designed. In this paper, our main focus is the visual feature extraction and the effect of these features to the recognition accuracy. We proposed an algorithm to extract visual speech features. The algorithm consists of two visual analysis stages: lip region detection and lip feature extraction. In the lip region detection stage, the speaker's mouth in the video sequence is located based on color and location information. The lip feature extraction stage extracts the lip region from its surroundings by utilizing both color and edge information.

In the lip feature extraction process, the key corners that define the lip region are detected and the relevant set of visual speech parameters is obtained. In our proposed visual front end processing, the lip geometric and motion features are extracted and were used as the input feature set to the recognizer. These features can improve the accuracy of the ASR system as compared with audio-only approaches.

In this paper, we will propose a fully automatic approach to extraction the visual features and investigate the performance improvement with the addition of visual feature for Mandarin audio-visual speech recognition as compared to audio only recognizer.

The organization of this paper is as follows. Section 2 presents our visual front end for lip feature extraction. The overall system and experimental results of speech recognition using visual and audio features are addressed in Sections 3 and 4, respectively. Finally, Section 5 discusses our conclusions.

II. AUDIO AND VISUAL FEATURES AND CLASSIFIER

A. Mandarin Speech Property

Mandarin is a monosyllabic language and some of its characteristics are significantly different with that of English or

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

other western languages. In Mandarin, the morpheme is formed by a sound plus a tone. Commonly used Mandarin characters are more than 10,000, all pronounced as monosyllables. A sound is the combination of an initial and a final in Mandarin. There are 21 initials and 16 finals. Tone is an important and distinct part of Mandarin. Each syllable of a Mandarin character needs to pronounce with a right tone or otherwise the meaning may be quite different. There are four tones plus one neutral. However, tones are probably the most difficult part of Mandarin. The combination of the initial, final and tone yields around 1,340 Mandarin sounds. In Mandarin, the initial is a consonant that begins the syllable while the final is a vowel that ends it. By monosyllabic property of Mandarin speech, we can consider that each Mandarin word utterance will have the same frame size if it is spoken at the same speed. Hence, we propose to use the distance-based classifier in the recognizer.

B. Acoustic Feature Extraction

According to the study of speech production, human can produce various sound by varying the shape of mouth and vocal tract. These properties are short-time stationary in terms of frequency domain response. In order to recognize a speech, we need to get features not only from time domain but also from the frequency domain.

The Mel-frequency cepstral coefficients (MFCC) have been shown to be more effective than other features in the speech recognition research. To obtain MFCCs, the input signal is preemphasized and divided into fixed length frames which may overlap with its adjacent frame to avoid the boundary effect. A hamming window function is applied to the frame before the short-time log-power spectrum is computed. Then the spectrum is smoothed by a bank of triangular filters, in which the pass-bands are laid out on a frequency scale known as Mel-frequency scale. The filtering is performed by using the DFT.

By the monosyllabic property of Mandarin as described previously, the number of samples of a Mandarin utterance can be viewed as fixed. Therefore, the MFCC of the frames in a word utterance forms the feature vector.

C. Visual Feature Extraction

The features for visual speech information extracted from image sequences can be geometric features, model based features, visual motion feature, and image based features[11], [13]. Geometric feature based techniques assume that certain measures such as height, width or area of the mouth opening are important. Petajan [25] developed an audio-visual recognition system that was based on geometric features of the mouth opening, like height, width and area. A simple threshold technique was first used to find the mouth area. Potamianos [13] and Zhi [10] described another semi-automatic lip-reading system which was based on the extraction of the lip contour features from the previous marked points. In its implementation, the speaker's lip contours were extracted from the image sequence. The motion-based feature approach assumes that visual motion during speech production contains relevant speech information. Visual motion information is likely to be robust to different skin reflectance and speakers.

D. Color Processing of Face Image

The RGB color model is widely used in computer vision because the red, green, and blue colors are the most natural color form and can be used to reconstruct almost all the visible colors. However, its inability to separate the luminance and chromatic components of a color hides the effectiveness of color in object recognition from image. To separate the luminance and chromatic components, various color space transform methodologies can be used, such as the normalized RGB space, HSL(hue, saturation, and luminance), HIS(hue, intensity, and saturation), HSV(hue, saturation, and value), and YCbCr.

The choice of an appropriate color space is an important factor for successful image segmentation and feature extraction. The measurement of skin reflectance, light spectral power distribution and camera channel sensitivities allow the computation of ideal RGB values for different skin types. The conversion to normalize color space (r, g) chromaticity is defined as

$$r = R / (R+G+B), \tag{1}$$

$$g = G / (R+G+B). \tag{2}$$

This transform can reduce the brightness dependency of skin color in an image. The conversion between RGB color and luminance is defined as

$$Y = 0.3\mathrm{R} + 0.59\mathrm{G} + 0.11\mathrm{B}. \tag{3}$$

In our automatic visual feature extraction operation, we use the normalized color space and the luminance.

E. Classifier

The nearest neighbor classifier can be used to improve the performance of the voice identification [1]. The WD-KNN classifier [6] is a weighted-distance classifier, derived from the basic unweighted-discrete KNN method. A short description of WD-KNN is given as follows. The collected features are split into data elements $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t$, t being the total number of training samples. The space of all possible data elements is defined as the input space $\boldsymbol{X}$. A feature map is defined to be a function that takes an input element in the input space and maps it to a point in the feature space. In general, we use ϕ to define a feature map and get

$$\phi : \boldsymbol{X} \to \boldsymbol{F} \tag{4}$$

When a test sample $\boldsymbol{y}$ and a specified distance measure are given, we obtain the k nearest neighbors belonging to class j, $N_{k,l}^{j}(\boldsymbol{y})$, which can be defined as:

$$N_{k,l}^{j}(\boldsymbol{y}) = \left\{ \boldsymbol{z} \in N_{k,l}(\boldsymbol{y}) : c(\boldsymbol{z}) = j \right\}, \tag{5}$$

where the cardinality of the set $N_{k,l}^{j}(\boldsymbol{y})$ is equal to k.

We now have k nearest neighbors of test sample $\boldsymbol{y}$ from each class. In the weighted-distance KNN classifier, we apply a weighting sequence to the selected neighbors in each class. Among the k nearest neighbors in class j, we define the weighting for each neighbor according to its distance to the test sample $\boldsymbol{y}$. We arrange the distance measure of the selected k neighbors as

$$dist_1^j \le dist_2^j \le \ldots \le dist_k^j \tag{6}$$

and define the w_i as the weight of the ith nearest neighbor. Since the one having the smallest distance value $dist_1^j$ is the most important, therefore, we set the constraint $w_1 \ge w_2 \ge \ldots \ge w_k$ when selecting the weights. The classification result $j^* \in \{1, ..., l\}$ is thus be

$$j^* = \mathbf{arg\,min}_{j=1,\ldots,l} \sum_{i=1}^{k} w_i dist_i^j \tag{7}$$

III. MANDARIN AUDIO-VISUAL RECOGNITION

The structure of an audio-visual recognition system is described as follow. The features for speech recognition are extracted from the video signal, including the acoustic corpus and image sequences. The extracted acoustic features from acoustic signal build up the audio feature. The visual feature can be obtained from the video. Then we combine the audio and visual features to establish the fusion features which are used to train and test in the recognition system. In the visual feature extraction stage, we extract the visual geometric and motion features of the lip using the method described above. The WD-KNN classifier is then used as the classifier in the recognition.

A. Face and Lip Region Detection

The lip region detection from the face image is to determine the bounding box for each organ. That is, we need to define a rectangular region which contains the region that we want to extract. The study in [16] reported that there are approximately 16 to 20 feature points or action units (AUs) need to construct the face organs. To find these AUs, we first divide the face region horizontally into two parts. The upper face region contains the eyes and the lower face region contains the nose and mouth. In the face detection step, our proposed face region detector is fairly accurate. So the relative size and position of the face box is consistent in all of our test images. Therefore, we can partition the face region horizontally into two equal halves. The upper face region is processed by edge detection following with the horizontal projection. The location of eyes can be identified from the peak of the projection histogram.

Most of the face region detection uses the color and luminance information to extract the face in the image [14-15]. We first transform the image into normalized color space. The luminance of the image is also obtained. We then build a mask from the transformed color space by threshold operation. According to our investigation, the value of r is between 0.3 and 0.6 and the value of g is between 0.2 and 0.4 for ordinary Asian face. After obtaining the color values of the r and g, the next step is to identify the value of the hair from the image by using the luminance of the image.

Next, the image is processed with the binarization operation according to the r, g and Y values to extract the face region from the image. A sample binary image and the segmented face region are shown in Fig. 1 and Fig. 2, respectively.

Figure 1. The binary image after skin color processing.

Figure 2. The segmented face region.

(a) Eye bounding box

(b) Nose bounding box

(b) Mouth bounding box

Figure 3. The detected bounding box of the (a) eyes, (b) nose, and (c)mouth.

To locate the middle point of the mouth we first define a bounding box for the mouth. Here we define the distance between the center points of the two eye as D_{eye}. Then, the nose region is the horizontal strip whose top and bottom edges are at $0.6D_{\text{eye}}$ and $0.85D_{\text{eye}}$, respectively, below the lower edge of the eyes has and with a width equal to $0.5D_{\text{eye}}$. The top of the mouth is the bottom of the nose. The bottom of mouth region is

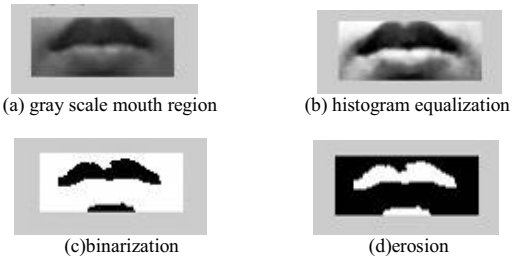
at $0.65D_{eye}$ from the bottom of nose region. Now, the bounding box for the mouth region can then be obtained. We can then compute the horizontal and vertical histogram of the binarized mouse image. From these two histograms, we can estimate the shape of the mouth. The center of the widest peak will define vertical position of the medial point of the mouth. By choosing the widest peak, the possibility of detecting the nose instead of the mouth is avoided. Fig 3(a) shows the detected bounding box of the eyes, while (b) and (c) illustrate the bounding boxes of the nose and mouth extracted from the face image, respectively.

B. Lip Feature Extraction

To split the lip region from the background, we use two threshold operations. The mouth image inside the bounding box is first transformed into the gray level image and processed by applying histogram equalization. Then a threshold is applied to the image which transforms the image into binary image. The gray scale image, equalized image and binary image of the mouth are shown in Fig. 4(a), (b) and (c), respectively. The erosion operation is then applied to the image to eliminate the unwanted noises. Based on the resulting image, we extract the lip region from its surroundings by finding the largest connected region.

Segmentation results from above processing steps are demonstrated in Fig. 5. The selected feature points are marked by small squares.. We observed that the feature points are all well match the true lip area. From the segmented lip image, we are able to extract the key feature points on the lips.

In our system, the bounding box is first extracted from the original image. Then we apply the proposed algorithm to automatically locate the six feature points on the mouth image. By locating the six feature points of the mouth for every video frame, we get the motion and geometric features for the recognizer. The main features, corners of the mouth, are then found as the cues for geometric parameters and motion vectors. In our system, the geometric parameters, including contour width, height, perimeter, area and motion feature vectors from subsequent frames are used as the feature vectors. Finally, these geometric parameters and motion vectors for selected feature points formed the feature vectors.



(a) gray scale mouth region (b) histogram equalization

(c)binarization (d)erosion

Figure 4. Process of mouth region image processing.

Figure 5. The extracted lip feature points used in the proposed system.

C. Audio Feature Extraction

The MFCC feature is the audio feature for our Mandarin audio-visual speech recognition system. The MFCC feature vector contains 12 cepstral coefficients extracted from the Mel-frequency spectrum of the frame with normalized log energy.

As described in Section II, we need to build the input feature vector for the WD-KNN classifier. The elements of the input space are mapped into points in a feature space $\boldsymbol{F}$. In our work, a feature space is a real vector space of dimension n, $\Re^n$. Hence, each point $\boldsymbol{f}_i$ in $\boldsymbol{F}$ is represented by an n-dimensional feature vector:

$$\boldsymbol{f}_i = (MFCC_{i1}, \ldots, MFCC_{ip}, VF_{i1}, ..., VF_{iq}) \quad (8)$$

where p, q are the dimension of MFCC and *Visual Feature*(*VF*), respectively, and

$$n = p + q. \quad (9)$$

By the monosyllabic sound property of Mandarin, the number of frames of a Mandarin utterance can be considered as fixed. The dimension p of MFCC feature is combined from all frames of a word utterance by average.

IV. Experimental Results

We applied the feature extraction algorithm on the audio-visual speech corpus. As the audio-visual database is concerned, there have been efforts in creating database for the audio-visual research area. Most of these databases are in English or other language, such as Tulips1, AVLetters, M2VTS [8], CUAVE, etc. Thus, to perform the audio-visual speech recognition in this research, we need to build an audio-visual database of Mandarin speech. The audio-visual database was recorded from 40 speakers. The video is in color with no visual aids given for lip or facial feature extraction. In this database, each individual speaker was asked to speak 40 isolated Mandarin digits, facing a DV camera.

In this experiment, the video stream is a sequence of 17 to 25 images for each utterance. Not all of the image sequences for Mandarin utterance were used in the recognition. In WD-KNN or distance-based classifiers, since the distance between the feature vectors is computed, the size of each feature vector must be the same. Visual features are selected from fixed number of images from each utterance.

The audio recognizer use 12 MFCC features extracted from speech sample at 8 kHz. The WD-KNN classifiers were implemented with k equals to 10.The selection of weight used in the WD-KNN is an important factor for the recognition

accuracy. In this paper, we use a reversed Fibonacci sequence as the weighting function in the WD-KNN classifiers. As discussed in [6], using reversed Fibonacci sequence can obtain the best classifier result. The reversed Fibonacci weighting function is defined as

$$w_i = w_{i+1} + w_{i+2}, \quad w_k = w_{k-1} = 1 \tag{10}$$

Table I summarizes the experimental results of different visual feature combination for each digit utterance under the 15 dB noisy condition. The accuracy ranges from 68.7%~88.7%, 64.4%~85.5% and 83.2%~94.3% in geometric-feature-only, motion-feature-only and geometric-motion conditions, respectively. It can be concluded that the recognition rate using geometric-only feature has better performance than that using motion-only feature in most of the cases. This implies that the geometric features seem to be more effective visual features than motion features for the speech recognition. The combined feature of the geometric and motion produces the highest correct rate among these three experiment conditions.

TABLE I. COMPARISON OF RECOGITION RATE USING DIFFERNET VISUAL FEATURES ON DIFFERENT DIGITS UNDER 15 DB NOISY CONDITION.

Digit words	Recognition rate (%)		
	Geometric feature	Motion features	Geometric + Motion
0	68	64	83
1	70	67	85
2	82	79	90
3	84	85	92
4	77	73	88
5	88	83	94
6	78	73	92
7	73	66	84
8	79	80	93
9	82	78	89

V. CONCLUSIONS

In this paper, our focus is to implement a recognition system using an automatic lip-geometric based feature extractor for the Mandarin audio-visual database with WD-KNN classifier. We develop a method of automatic lip feature extraction and its application to Mandarin audio-visual speech recognition. Our algorithm first reliably segments and locates the mouth region by using normalized RGB space from a color video sequence. The algorithm subsequently segments the lip from background by making use of both color and edge information. The lip key points that define the lip position are detected and the relevant visual speech parameters are derived and form the input to the recognition engine. We then demonstrated three experiments by exploring these visual parameters.

The results are compared with different visual features. Comparison among these visual features for Mandarin speech recognition and the improvement of using both geometric and motion visual feathers are shown and discussed. It was found that by enabling extraction of an expanded set of visual speech features including the geometric and the motion features of the lip, the proposed visual front end achieves an increased accuracy when compared with previous studies that use only lip geometric features.

REFERENCES

[1] L.G. Bahler, J.E. Porter and A.L. Higgins, "Improved voice identification using a nearest-neighbor distance measure," *Proc. Acoustics, Speech, and Signal Processing*, Vol. 1, Issue:19-22, pp. 321-323, Apr. 1994.

[2] T. Chen, "Audio-visual speech processing," *IEEE Signal Processing Magazine*, Jan. 2001.

[3] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23-37, Feb. 2002.

[4] D. DeCarlo and D. Metaxas, "Optical Flow Constraints on Deformable Models with Applications to Face Tracking," *Int'l J. Computer Vision*, vol. 38, no. 2, pp. 99-127, 2000.

[5] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," in *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, Sept. 2000.

[6] Tsang-Long Pao, Yu-Te Chen and Jun-Heng Yeh, "Comparison of classification methods for detecting emotion from Mandarin speech," *IEICE Transactions on Information and Systems*, Vol. E91-D, No. 4, pp. 1074-1081, Apr. 2008.

[7] M.I. Faraj and J. Bigun, "Person Verification by Lip-Motion," *Proc. Conf. Computer Vision and Pattern Recognition Workshop* (CVPRW '06), pp. 37-45, 2006.

[8] M.I. Faraj and J. Bigun, "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition", *IEEE Trans. Computers, ,* vol. 56, no. 9, pp.1169 - 1175, Sep. 2007.

[9] M. Heckmann, F. Berthommier and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal processing*, 2002:11, pp.1260-1273, 2002.

[10] M. N. Kaynak, Q. Zhi, etc, "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition," *IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, pp. 564-570, July 2004.

[11] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. pattern analysis and machine intelligence*, vol. 24, pp. 198-213, 2002.

[12] E. D. Petajan, N. M. Brooke, B. J. Bischoff, and D. A. Boddoff, "Experiments in automatic visual speech recognition," in *Proc. 7th FASE Symp.*, Book 4, pp. 1163-1170, 1988.

[13] G. Poamianos, etc,"Recent Advances in the Automatic Recognition of Audiovisual Speech" *Proceeding of the IEEE*, vol. 91, no. 9, Sep. 2003.

[14] L. Sigal, S. Sclaroff and V. Athitsos, "Skin Color-Based Video Segmentation under Time-Varying Illumination", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, Issue: 7, pp. 862-877, July 2004.

[15] M. Soriano, B. Martinkauppi, S. Huovinen and M. Laaksonen, "Skin detection in video under changing illumination conditions", *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 1, pp. 839 – 842, 3-7 Sept, 2000.

[16] M. Valsta and M. Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis", *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop* (CVPRW'06), pp. 149 – 149.