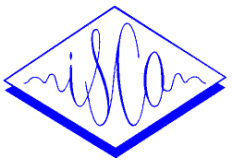

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

TOWARTDS ROBUST LIPREADING

Wen Gao[1,2], *Jiyong Ma*[1,2]

[1]Institute of Computing Technology
Chinese Academy of Sciences
Beijing 100080, China

Rui Wang[2] *and Hongxun Yao*[2]

[2]Department of Computer Science
Harbin Institute of Technology
Harbin 150001, China

ABSTRACT

In this paper, a robust and fast approach to lip detecting and lip-reading is presented. The approach combines the information of lip color with the geometrical features of lips in human face. This technique makes it possible to derive lip regions in real time under regular illumination conditions. The experimental results with more than 2000 images have shown that the approach to locating lips is very efficient both in locating speed and locating accuracy. Recognition tests were conducted on Chinese phrases. The approach achieved an accuracy of 90% for speaker dependent recognition task.

1. INTRODUCTION

Automatic lipreading is complementary to both speech recognition and hand gesture recognition. However, it is difficult to integrate the visual speech with auditory speech. The main task in incorporating the visual information into a speech recognition system and hand gesture recognition is to find a robust and accurate approach for tracking the lip movements in real time and extracting important features. It has been widely accepted that the information of lip motions is not enough for lipreading [1]. It also requires the visual information of the mouth. Several automatic lipreading systems were developed in the past few years. Chiou GI[2] designed and implemented a lipreading system that recognized isolated words using only color video of human lips. The system performed video recognition using "snakes" to extract visual features of geometric space, Karhunen-Loeve transform (KLT) was used to extract principal components in the color eigenspace, and hidden Markov models (HMM's) were used to recognize the combined visual features sequences. With the visual information alone, the system achieved 94% accuracy for ten isolated words. Silsbee PL.[3] developed a lipreading system in conjunction with an audio ASR system in order to improve the accuracy of the latter, especially under degraded acoustical conditions. Experimental results were presented for two small phoneme discrimination tasks, as well as a medium vocabulary isolated word recognition task. In all cases, performance of the combined system was superior to that of the audio system, with a reduction in errors ranging from 20 to 65%.

To date, there are two kinds of approaches to feature extraction: the model-based approach and the image-based approach. For the case of model-based approach, a geometrical model of the lip contours is applied to the input image of the speaker' s lip. The typical examples were Rabi G[4] and Luettin J[5] described a robust method for extracting visual speech information from the shape of lips to be used for an automatic lipreading systems. The disadvantage of this kind of approach is time consuming during the match of the lip model.

For the case of the image-based approach, the interest region is segmented. The advantage of this kind of approach is its fast speed. Recently, Gray, Movellan and Sejnowski [6] compared a wide variety of images processing techniques, sensor fusion, and recognition engines for automatic lip-reading. The basic findings were that: (1) It is a good idea to use texture information, not just lip contours; (2) global image decomposition (e.g., PCA) does not work as well as local representations (e.g., Gabor wavelets); (3) As in the acoustic domain the best sequence recognition engines are HMMs.

Mase K. and Pentland [7]proposed an approach for automatic lipreading by optical-flow analysis to against illuminant variations.

Lip tracking also has many applications such as in lip motion synthesis, and in Synthetic/Natural Hybrid Coding (SNHC) in MPEG4, multimedia communication and speech-driven talking heads [8,9,10]. Massaro and his colleagues [11] demonstrated that human perceivers combine acoustic and visual information in a factorized manner, i.e., as if acoustic and visual sources were class conditionally independent. This fact greatly simplifies the integration problem, since it suggests that separate acoustic and visual recognition engines may be developed and then combined under the assumption of conditional independence. Movellan and Mineiro [12] pointed out the relative weight of the Audio and Visual recognition engines must be adjusted on-line; how to weight these sources automatically during system use to optimize recognition performance is an important research challenge.

For robust lipreading, Tanaka A [13] proposed an approach to robust lipreading by using intensity and location normalization. Experimental results showed that the recognition rates had been very much improved by the normalization techniques. Uwe

Meier[14] discussed robust lipreading. The problem of fast and robust locating lips is still a challenge because the lip contours of different people varies greatly and are affected by lip variations and head movement and illuminant conditions while a speaker is uttering. To solve this problem, many approaches were proposed to describe the lip shape such as deformable template [15], snake [16] and active shape models [17] and [18] etc. Among them, deformable template is widely accepted due to its advantages of simplicity, easy use for locating lips, etc. The grayscale image was used in early research work. Because the difference between grayscale information of lip color and that of skin color is small, the accuracy of detecting and locating lips is not high, and the requirement of illuminant conditions is rigorous. Experimental results have shown that lip color information is more robust to lighting conditions. With the increase of computer speed, colors of human face skin and lips were paid more and more attention , colors provide richer information than grayscale , it is possible to increase the accuracy of lip locating with color information. One key technique is the transform from the RGB color images to other color spaces such as YUV space, HSV space, YIQ space and rgb space, etc. After the transform the intensity information is omitted. One component or two components, which have more discriminant power to separate the skin color and lip color, are selected and used to detect and locate lips. The approach using linear-scale range of lip color to locate lip during is not accurate. The approach based on the probability distributions of lip color and skin color was proposed to increase the accuracy of detecting and locating lips. Because the lip region must be in a face image, the decision about if a region is a lip region becomes a two-class discriminant problem, namely, face or lip. The Bayessian decision is usually used. However, the threshold setting becomes a key problem, a bigger threshold makes more false reject rates and a smaller one makes more false acceptance rates. Although color information indeed reduces the illuminant effects, it can not completely exclude the illuminant effects and the individual variations. Therefore, the locating approach has no adaptivity

Fisher transform is a kind of linear discriminant transform. It uses the principle of maximizing the ratio of inter-class distance to within class distance and transforms the multi-dimensional space to one-dimensional space. The approach can not only enhance the discrimination between lip color and skin color but also enhance the lip contours. Therefore, it increases the accuracy of locating lips. And this approach uses the lip color and skin color discriminant information, this provides a degree of invariance to changes in illummination. Although the relative distributions between lip color and skin color are considered in the approach, the locating approach has no adaptivity either. Therefore, other approaches need to be developed to realize adaptively locating lips.

In summary, the Fisher's discriminant analysis itself only considers the distributions of lip color and skin color, and doesn't consider the geometrical distribution of lip contours in a face. It will be advantageous to lip locating using the geometrical distribution of lip contours in a face. The conventional approaches to lip locating don't solve the problem of locating inner lips. Because the inner lip is subject to the effects of teeth, black hole and tongue, this enables locating inner lips inaccurately. Although, some systems utilized the penalty function to increase the accuracy of locating inner lips, this problem of accurately locating inner lips is not completely solved. Resolving the accurately locating inner lips becomes a challenging.

Fisher transform with constraints is proposed to enhance the lip region of a face image. Two techniques are proposed to facilitate locating lips. One is the relative area of the lip in color to that of a face region is almost invariant to a specific person, this characteristic is used to adaptively set threshold to distinguish the lip color and skin color. The other is the relative area of upper lip or lower lip to that of the square whose length equals to the distance of two lip corners is almost invariant. A linear correlation between parameters of inner lip and those of outer lip can be obtained based on this assumption. This linear correlation is used to predict parameters of the inner lip by parameters of the outer lip. The local minimums of cost function can be overcome by using this technique. And the locating accuracy is increased.

The organization of the paper is as the following: Section 2 describes the Fisher transform with constraints. Section 3 discusses the robust lip tracking approach. Section 4 describes the feature extraction approach. Section 5 describes the recognition approach. Section 6 demonstrates the experimental results. Section 7 contains conclusion.

2. FISHER TRANSFORM WITH CONSTRAINTS

Fisher's discriminant analysis is a kind of approach to distinguishing between lips and skin. Its basic idea is to transform the original data to maximize the separability of the two classes. The RGB space is transformed into YIQ space. Two components Q and phase angle FI are set to a vector x which used in Fisher transform to distinguish the skin color and lip color. Because there are other color classes in a face besides the skin color and the lip color, for example, the eye color, eyebrow color, nostril black color, etc, and the colors of these parts are very near the lip color after fisher transform. This affects on the accuracy of lip detecting. Therefore, some constraints must be developed. The colors of pixels not occupied in the skin area and the lip area are replaced with the average skin color. This is achieved by choosing the colors of pixels, which colors are far from the lip color and skin color. The Fisher transform with the constraints is as the following

$$y = \begin{cases} x^T W & \text{if } x \in \boldsymbol{e}_{lip} \text{ or } x \in \boldsymbol{e}_{face} \\ m_1^T W & \text{else} \end{cases}$$

Where y is the image after Fisher transform, $\boldsymbol{e}_{lip}$ denotes the set of lip color, $\boldsymbol{e}_{face}$ denotes the set of face color. The Fisher transform with constraints reduces the effects of other components in a human face on lip locating. The images transformed by the Fisher transform are shown in Fig.1. The lip color is enhanced. The skin and lip colors are separated.

3. ROBUST LIP TRACKING

3.1 Lip Segmentation

One fact is the relative area of the lip in color to that of a face region is almost invariant to a specific person, this characteristic is used to adaptively set threshold to distinguish the lip color and skin color. Note that the lip area is enhanced after Fisher transform, i.e. the lip area is the brightest area in a face. The threshold to separate the lip area and skin area can be determined according to the above observation and statistics. The threshold is used to binarize the enhanced face image. Suppose that the color distributions of skin colors and lip colors are Gaussians. The mean and standard deviation of lip colors are $\boldsymbol{m}$ and $\boldsymbol{s}$ respectively. The ratio of lip area to skin area is within 4%-7% according to our statistic. Although the ratio is different for different people, but its change range is small. The ratio value is set to 5% in our system, the corresponding threshold is $m + 1.65s$ according to statistics. As shown in Fig.1, the lip area can be detected for a face under poorer illuminant condition and lip states. The detecting accuracy of regions of interest (ROI) is 100%. This approach to setting threshold has very good adaptivity for different environments. This approach not only can increase the accuracy of locating lips but also has a very fast speed. It can achieve real-time implementation. The outer lips are easily found using this approach, including the key points of outer lips such as the top point of upper outer lip and the lowest point of the outer lower lip.

(a) (b) (c) (d)

Figure1. (a)The original color face image and the lip region. (b) The enhanced face image using lip color information. (c) The binarized face image. (d) The image after morphological opening

3.2 Lip fitting by Deformable Template

The initial template is defined as shown in Fig.2. Adapting the template to match the lip contour involves changing the parameters to minimize cost function. The cost function in our system includes information of lip edges, space constraints, etc.

3.2.1 Potential fields

The lip contours are main information for template locating. Since the edges are mostly horizontal, the vertical gradient of the image is used as the potential field, i.e. 3×3 Prewitt operator is used to enhance the edges of the image after Fisher transform. The advantages of the Prewitt operator are not only it is sensitive to horizontal edges, but also it differentiates between "positive" and "negative" edges. A positive edge is one where the image intensity is higher above the edge than below it, and vice versa. A template that uses this difference will not confuse the upper and the lower lip. In order to ensure the continuity of the cost function, the Gaussian operator is used to smooth the image.

3.2.2 Cost function

The deformable template is used to describe the lips and locate the lips. As shown in Fig.2, the outer lips are described with two quartics and the inner lips are described with two parabolic curves. The cost function consists of three terms, i.e. potential fields, integrals and penalty terms [19].

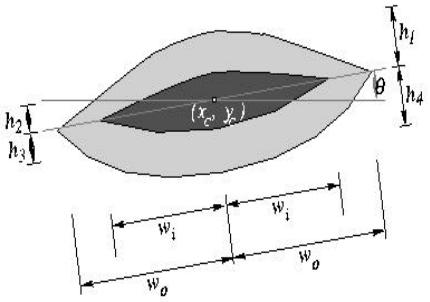

Figure 2.The Deformable Template and Its Associated Curves

The cost function includes four curve integrals, one for each of the four lip edges. The cost function also includes a number of penalty terms. The constraints make sure that the parameters of the template stay within reasonable limits. For example, the constraint $h_1 > h_2$ makes sure that the inner lip is not higher than that of the outer lip, the constraint $a_1 < (h_1 - h_2) < a_2$ makes sure that the thickness of lips is within reasonable limits. If the constraint is satisfied, the cost function will increase quickly. In addition, there are other penalty terms which make the template is within reasonable limits.

The outer lip contours becomes clearly after Fisher transform. And the initial position of the template obtained by the adaptive threshold setting approach discussed in section 3.1. Therefore, the outer lip contours are easily matched by using the deformable template approach discussed above. As shown in table 1 that the accuracy of locating the outer lips can reach as high as 96%, no match error occurs. But the inner lip is

subject to the effects of teeth, black hole and tongue, this enables locating inner lips inaccurately. To solve this problem, the initial parameters of the inner lips need to be set. The linear correlation between outer lips and inner lips is used to predicate parameters of the inner lip by parameters of the outer lip.

3.2.3 Predicting the positions of inner lips using multivariate regression

It is assumed that the height (h2) of the inner lip has a linear relation with the parameters (W_0, h1、 q1) of the outer lip. Similarly, it is assumed that the height (h3) of the inner lip has a linear relation with the parameters (W0、 q2、 h4) of the outer lip. Therefore, we have the following equations

$$h_2 = A_1 h_1 + B_1 q_1 + C_1 W_0$$

$$h_3 = A_2 h_4 + B_2 q_2 + C_2 W_0$$

The multivariate regression is used to estimate the parameters in the above equations. The parameters of the deformable templates of ten persons are used to estimate the parameters.

To verify the linear assumption, the test results of predicted inner lip parameters h2、 h3 with 200 lips are shown in the Fig.3. The bold line is the true value and the thin line is the predicted value. From the figure, it is seen that the predicted value is very near the true value. Therefore, the assumption is hold.

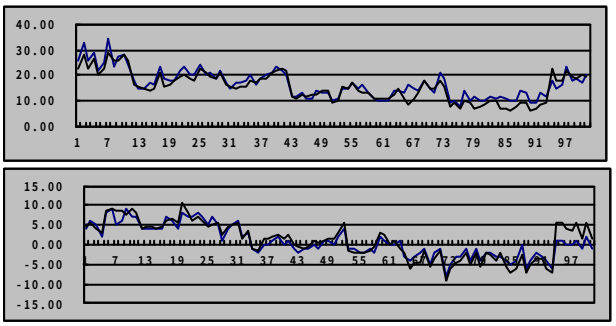

Figure 3. The comparison between the actual values and predicted values of the inner lip parameters. The upper one is for the parameter h2, and the lower one is for h3

The above-obtained linear relationship of h2 and h3 of inner lip parameters are used as the initial value for deformable template. This increases the accuracy of inner lip locating and avoids the issue of local minimum. As shown in Fig.4, the original images around the lips are shown at the top. The lip locating result without using linear prediction is shown at the middle, the inner lip locating is not accurate. The lip locating result using linear prediction is shown at the bottom, the locating result is very accurate. The locating speed is also increased using the predicted parameters of inner lips.

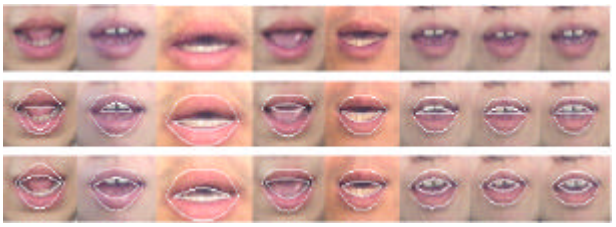

Figure 4. The results obtained by the linear prediction

4. FEATURE EXTRACTION

Subsequent processing is restricted to the segmented lip area. The original RGB image is converted into an intensity image. For illumination invariance, each pixel intensity value of is normalized by dividing the average intensity value of the segmented lip area. The segmented area is further segmented into 5*5 sub areas. The average intensity value of each sub area is used as feature. They consist of a vector with 25 dimension..

5. RECOGNITION APPROACH

Hidden Markov Models (HMMs)[20] have been used successfully in continuous speech recognition, handwriting recognition, etc. An HMM is a doubly stochastic state machine that has a Markov distribution associated with the transitions across various state, and a probability density function that models the output for every state. In the case of lipreading, we used phrase HMMs for each phrase to be recognized. The models are trained used the Baum-Welch algorithm. Viterbi algorithm, which computes the likelihood for each HMM of having generated the observed sequence and the model with the highest likelihood, is chosen as the recognized phrase. The structure of HMM is left to right, the number of states of each phrase is 10.

6. EXPERIMENTS

6.1 Lip-tracking

Experiments were conducted to compare the accuracy of detecting regions of interest (ROI) and lips with linear prediction and without linear prediction. The experiment environment:CPE-3000 image acquisition card, JVC TK-1070 color camera and MIMTRON MTV-33-1 CB color camera, the frame rate is 25 frames per second, the CPU of computer is Pentium-II 300.To show the advantages of the approach proposed in this paper, experiments were conducted using the active shape model and the Fisher transform. The comparison

between them were conducted by using the image enhancement approach with Q, FI components. The image database consists of images from 4 males and 5 females, the total images are 2000, which were collected under different illuminant conditions and different utterances. The subjective evaluation approach to locating lips [4] was used. The results of this approach are good, Adequate and Miss. The approach is an important model for evaluating different approaches for locating lip contours in robustness and accuracy. A locating result was classified as good if the lip contour was found within about one quarter of the lip thickness deviation. It was classified as Adequate if the outline of the contour was found between one quarter and half the lip thickness deviation and it was classified as a Miss otherwise. The average locating time is the sum of detecting and locating time. The detection means the ROI of lips is completed detected and the error range is within one quarter the lip deviation. In the table 1,the first row of the result is the detecting result without using predicting, instead the second row is shown the detecting result using predicting. The ROI detection speed is 18-20 frame per second. Average locating lip contours time is 6-8fram/s. And the outer lip locating accuracy is 96%.

Table 1. Comparison among different approaches

Inner lip locating			Average locating Time
Good	Adequate	Miss	
70%	5%	25%	6-8fram/s
84%	13%	3%	6-8fram/s

6.2 Lip-reading

Experiments were performed using a database, which consists of color image sequences of ten phrases. Each phrase was spoken 5 times by a male speaker. Four was used for training, one for test. Each image sequence consists of twenty frames. The ten Chinese phrases (two words) not correctly recognized by sign language recognizer are ǰueqi/, ṣhengchan/, ḳuoda/, /laoban/,/zhuanshun/,/sikao /jisuan/, /kunnan/,/koudai/,/kaifa/. Among the ten phrases, only one word was not recognized correctly. If only lip shape such as lip width or height was used as feature, no phrase can be recognized correctly, this result shows that the lip shape has little discriminant power to lip-reading. In fact, the lips, teeth, tongue and lip shape all affect the recognition performance. The tracking result is shown in the Fig.5.

Figure 5. The tracking results of the Chinese phrase /jueqi/

From the table, it is seen that the outer contours of lips are located ver accurately using the approach proposed in this paper, no error occurs. Both the speed of detecting ROI and the speed of locating lips are very fast. The approach using linear prediction increases the accuracy of locating inner lips, however, there are still 3% errors. These errors mainly occur when the phoneme /u/ is uttered. Because the lips contracted tightly when the phoneme /u/ is uttered, the linear relation between parameters of inner lips and parameters of outer lips is not hold. And the inner lips become a small area like a point. To locate this type of inner lips, special approach needs to be developed.

Experiments were performed using a database that consists of color image sequences of ten phrases. Each phrase was spoken 5 times by a male speaker. Each image sequence consists of twenty frames. Four is used for training, one for test. In the experiments for locating and tracking lips, only five frames are located mistakes. Among the ten words, only one word is not recognized correctly.

7. CONCLUSION

We have developed a novel approach to real time automatically tracking lip regions and have applied this approach for lipreading. Experimental results show that this technique is capable of improving both the recognition performance and speed. The approach for tracking lip regions is robust and fast.

References

[1] R.Kaucic,B.Dalton,and A.Blake.Real time lip-tracking for audio-visual speech recognition applications. In Fourth European Conference on Computer Vision, Vol2 ,pages 376-386.Cambrige, 1996.

[2] Chiou GI. Jenq-Neng Hwang. Lipreading from color video. IEEE Transactions of Image Processing, vol.6, no.8, Aug. 1997, pp.1192-5. Publisher: IEEE, USA.

[3] Silsbee PL. Bovik AC. Computer lipreading for improved accuracy in automatic speech recognition. IEEE Transactions on Speech & Audio Processing, vol.4, no.5, Sept. 1996, pp.337-51. Publisher: IEEE, USA.

[4] Rabi G. Si Wei Lu. Energy minimization for extracting mouth curves in a facial image. Proceedings. Intelligent Information Systems. IIS'97

[5] Luettin J. Thacker NA. Beet SW. Locating and tracking facial speech features. Proceedings of the 13th International Conference on Pattern Recognition. IEEE Comput. Soc. Press. Part vol.1, 1996, pp.652-6 vol.1. Los Alamitos, CA, USA.

[6] Gray, M. S., Movellan, J., and Sejnowski, T. Dynamic features for visual speechreading: A systematic comparison. Advances in Neural Information Processing Systems, 9. MIT Press, 1997.

[7] Mase K. Pentland A. Automatic lipreading by optical-flow analysis. Systems & Computers in Japan, vol.22, no.6, 1991, pp.67-76. USA.

[8] Rao RR. Tsuhan Chen. Mersereau RM. Audio-to-visual conversion for multimedia communication. IEEE Transactions on Industrial Electronics, vol.45, no.1, Feb. 1998, pp.15-22. USA.

[9] Lavagetto F. Converting speech into lip movements: a multimedia telephone for hard of hearing people. IEEE Transactions on Rehabilitation Engineering, vol.3, no.1, March 1995, pp.90-102. USA.

[10] Chen T. Hsieh H-Y. Lee L-S. Lip synchronization for Mandarin speech: A marriage of speech and image processing in the Mandarin environment.1995 International Symposium on Communications. Nat. Taiwan Univ. Part vol.1, 1995, pp.348-55 vol.1. Taipei, Taiwan.

[11] Massaro, D. and Stork, D. Speech recognition and sensory integration American Scientist, May 1998.

[12] Movellan J. R. and Mineiro, P.: Robust sensor fusion: Analysis and application to audiovisual speech recognition. Machine Learning, 32,85-100, 1998.

[13] Tanaka A. Vanegas O. Tokuda K. Kitamura T. Intensity /location normalization for automatic lipreading. 1998 Fourth International Conference on Signal Processing. IEEE. Part vol.2, 1998, pp.920-3 vol.2. Piscataway, NJ, USA.

[14] Uwe Meier,Rainer Stifelhagen,Jie Yang and Alex Waibel. Towards unrestricted lip reading.ICMI,1999

[15] A. L. Yuille, P. Hallinan, D.S.Cohen, Feature extraction from faces using deformable templates. Int. Journal of Computer Vision,Vol.8(2),pp.99-112,August 1992.

[16] M.Kass,A.Witkin,and D.Terzopoulos. Snakes: Active contour models. Int. Journal of Computer Vision,pp.321-331,1988

[17] J.Luettin,Neil A.Thacker,S.W.Beet. Locating and Tracking Facial Speech Features. International Conference on Pattern Recognition. Vienna, Austral, 1996.

[18] R.Kaucic,B.Dalton,and A.Blake.Real time lip-tracking for audio-visual speech recognition applications. In Fourth European Cofference on Computer Vision , Vol2 ,pages 376-386. Cambridge, 1996

[19] M.E.Hennecke,K.V.Prasad,D.G.Stork, Using Deformable Templates to Infer Visual Speech Dynamics. In 28th Annual Asilomar Conference on signals, Systems and Computers. IEEE, November 1994

[20] Rabiner, L. R. and Juang, B.H. Fundamentals of speech recognition, Prentice Hall, Englewood Cliffs, 1993.