行政院國家科學委員會專題研究計畫　成果報告

視覺化中文讀唇系統之研究

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-041-002-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：嘉南藥理科技大學資訊管理系

計畫主持人：黃文楨

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 9 月 24 日

A Mandarin lip-reading System (NSC 92-2213-E-041-002)

Wen-Chen Huang(黃文楨)

Department of Information Management, National Kaohsiung First University of Science & Tech.

Email: wenh@ms2.hinet.net

ABSTRACT

The most common communication disorders clinically are speech and hearing impairment. With the lengthening of life expectancy, hearing impairment has become one of the most important issues in our society. After medical or surgical therapeutic processes, rehabilitation is the last resort to restore the capability of communication. For speech and hearing disorders, communication can usually be augmented by visual stimulation, for example, by signs or text. Owing to the development of the world-wide-web, not only can our traditional oro-aural communication mode be replaced by cyber-text and distant communication, but the communication mode of the speech- or hearing-impaired people can also be substituted. For the proposed techniques of data analysis, we can also build a visual training tool for these patients, either for their post-operation communication or for their later speech rehabilitation period. The purpose of this research is to build a visual training tool which recognizes the sentence from the image sequences of speakers. At the same time, many face images are taken from different angles by digital cameras to build a vivid 3D human head model. The recognized text drives the 3D human head model to talk like a real person. There are three parts in our framework: visual speech recognizer, 3D head model generator, and talking face animation. The integrated system is combined with the web-based visual communication interface. Experimental results show that the recognition rate is about 97 percent of 10 sentences for a specific person who speaks Mandarin.

1. INTRODUCTION

The development of visual communication tools for hearing- or speech-impaired patients is made possible due to advances in computer graphics and computer vision. Many successful results can be found in different applications such as speech training, e-learning, tele-conference, and long-distance medical care. The purpose of this research is to build a visual speech recognition system which recognizes the Mandarin sentence from the image sequences of speakers. At the same time, many face images are taken from different angles by digital cameras to build a life-like 3D human head model. The recognized text drives the 3D human head model to talk like a real normal person. With this technology, speech- or hearing-impaired patients can communicate and stay in contact through the internet. Figure 1.1 shows the framework of the proposed method. There are three parts in our framework: visual speech recognizer, 3D head model generator, and talking face animation. The assumption of this system is sentence-based recognition on limited vocabulary for specific patients.

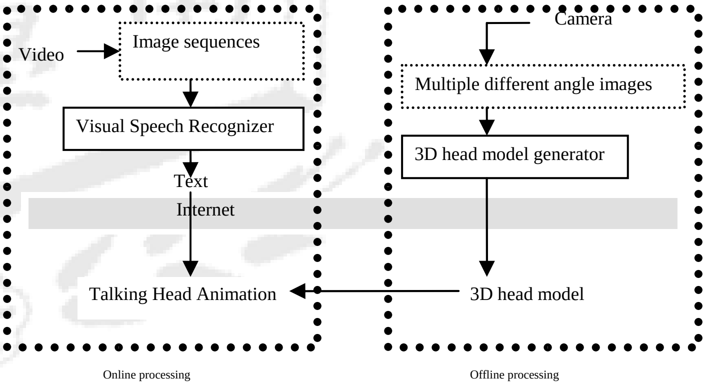

Figure 1.1: Framework Overview

For the visual speech recognizer, a spatial-temporal image differences (STID) method is used to recognize limited Mandarin sentences from the same speaker without voice. The image sequences of multiple sentences are the training and testing data. Each sequence of images of a sentence is a basic unit for the training or testing data. The sequences of images include the entire face region with complex background (not limited only to the lip region as most visual speech systems nowadays) under normal illumination. The registration and crop process are executed through all the image sequences. To remove the complex background, frame difference accumulation is used to distinguish the apparent moving part including the lip movement. The lip region is located by using the (Hue, Saturation, Intensity) HSI color model. After finding the lip region, the center of lip region is easily calculated by averaging the x-y coordinate of all the pixels upon the lip region. For feature extraction, we propose using the spatial and temporal features from image sequences. The temporal difference image makes the recognizing scheme more robust. The Euclidean Difference method is used for recognition.

The contribution of this research is to propose an integral visual education system for those patients who cannot speak out with a voice or have the difficulties hearing. The visual talking education tool is developed for speech- or hearing-impaired patients by showing the movement of lip to correct their pronunciation. The proposed system combines the recognition of Mandarin-speaking lip-movement scheme and 3D animation functions with the remote communication capability through the internet environment. The capability of correcting the sentence spoken by the speech- or hearing-impaired patients with more precise sentence is also demonstrated. Unlike other avatar systems, a new web-based visual communication interface is proposed. The advantage of our recognition method is simple and robust. With this system, speech- or hearing-impaired patients would be able to have visual communication as they meet the real persons in remote position.

The rest of the paper is organized as follows. Section 2 introduces previous related works. The research methods describing how to build the visual speech recognition system and the detailed algorithms are explained in Section 3. In Section 4, we describe the 3D head model generation method and also explain the talking face animation procedure in our system. Section 5 illustrates the experimental results. Finally, the conclusions and future works are given.

2. RELATED WORKS

The bimodal audio-visual system has recently attracted more attention in the speech recognition research field. Matthews, et al. [3] reported methods of extracting visual features for visual speech recognition. Three methods for parameterizing lip image sequences for recognition using hidden markov models are compared. Silsbee and Bovik [4] conducted a survey of the computer visual speech research. In visual speech recognition, there are two kinds of approaches, contour-based and image-based, to extracting features from a lip image. The contour-based approaches obtain features from lip contours using edge information [5], deformable templates [6], or active contour [7]. The advantage of contour-based features is invariant under spatial translation, rotation, scaling, and illumination. However, important information, such as teeth and tongue, is lost for the contour-based features [8]. The feature extraction of image-based approaches involves transformations with very little information reduction, such as principal component analysis [9], wavelet [10], and Fast Fourier Transform [11].

There are various techniques developed for 3D human face reconstruction. For example, laser scanner is used to get 3D coordinate value of object and build 3D model [12]; structured light is used to generate fixed stripes on the object to get 3D coordinate value [13]; another method uses both frontal face and side-face to build 3D model, and usually these two images are taken on orthogonal basis [14][15] ; using the corresponding point matching technique for computing two 3D coordinate value is exploited to build a 3D face model [16]; and image sequences are employed to reconstruct the 3D face model [17][18]. Among the above methods, some need special devices, while others need a lot of manual adjustment.

For talking-head animation, Chandrasiri et al. [19] presented a system for internet communications that combines text chatting with a real-time 3D facial expression synthesis – a talking head agent. Tao et al. [20] addressed the compression issue of MPEG-4 facial animation parameters for transmission of talking head. Life-like talking face animation with visual text-to-speech synthesizer is developed by Cosatto et al. [21] Chuang and Bregler [22] proposed a facial animation method using a combination of motion capture data and blendshape interpolation.

Our work is a hybrid method combining model-based and sample-based animation with recognition scheme. Text is generated after the visual speech recognition system instead of typing by the user.

3. Visual speech recognizer

The proposed system as shown in Figure 3.1 consists of six components: (1) Registration of image sequences, (2) Face region detection, (3) Lip region detection, (4) Spatial-temporal sampling of image sequences, (5) Training the resulted image sequences, and (6) Recognition .

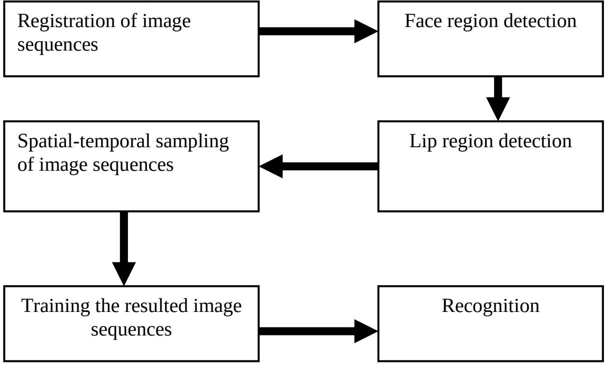
Registration of image sequences
Face region detection
Lip region detection
Spatial-temporal sampling of image sequences
Training the resulted image sequences
Recognition

Figure 3.1: Information flow of the proposed visual speech system

3.1 Training image sequences

The sample image sequences were digitized at 30 fps, 24-bit per pixel (RGB) format. The image size of each image is 352 by 240 pixels. Each sentence was recorded 10 times as the training image sequences.

3.2 Registration of image sequences

For registration of image sequences, we implement multiscale affine and rigid (rotation and translation) image registration algorithm [23]. An affine flow field can be expressed as:

$$\vec{\theta}(x,y) = \mathrm{A}(x,y)\vec{\rho} \qquad (3.1)$$

where $\vec{\rho} = (\rho_{1:}\rho_{2:}\rho_{3:}\rho_{4:}\rho_{5:}\rho_6)^T$ are the affine parameters to be estimated and

$$\mathrm{A}(x,y) = \begin{pmatrix} 1 & x-x_0 & y-y_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_0 & y-y_0 \end{pmatrix} \qquad (3.2).$$

All the sequences of images from the training and test database are required to be registered as the same base reference image such that all the images are in the same position.

3.3 Lip Region Detection:

The motion of a face provides useful information for detecting the face region even after image registration. Frame difference accumulation is used to distinguish the apparent moving part including the lip movement. Figure 3.2(a) shows the summation of frame differences displayed in gray level. Figure 3.2(b) shows the binary image after global threshold [1].

The portions of eyes, mouth, eyebrow, nose, and a few boundaries of head are segmented. In order to have clear visualization, the detected features with green color are overlapped onto original image and are shown in Figure 3.3(b). For those detected figures, lips have apparent visual color. As noted earlier, the RGB color model is not well suited for describing colors for human interpretation. In other words, the HSI model is an ideal tool for color descriptions that are intuitive to humans. Therefore, the transformation from RGB model to HSI model is performed. The result is shown in Figure 3.3(c). The mouth portion has distinct feature from other regions. Therefore, the lip region is easily detected and is shown in Figure 3.3(d).

After finding the lip region, the center of lip region is easily calculated by averaging the x-y coordinate of all the pixels upon lip region. The registration and crop processes are executed through all the image sequences. Figure 3.4 shows some examples of cropped image sequences.

3.4 Spatial-temporal sampling of image sequences

The overall image sequences are denoted as I_{rpq}, where the sentence: r=1,..,M, training times: p=1,..,N, and the decomposed image sequences: q=1,..,K$_p$. In general, M and N are fixed, but K$_p$ is different in each case (please see Table 1). The purpose of spatial-temporal sampling of image sequences is to adjust K$_p$ into a fixed number K. For every r and p, there are image sequences I_{rp1}, I_{rp2} ,.., I_{rpt} where t is the number of image sequences, say K$_p$. The difference in image sequences is

$$d_i = \| I_{rpi} - I_{rp(i+1)} \|, \; i=1, .., t-1. \qquad (3.3)$$

Find the minimal index of d$_i$, say b. In other word,

$$b = \arg(\min_i(d_i)) \qquad (3.4)$$

where i=1, .., t-1. Let

$$I_{temp} = (I_{rpb} + I_{rp(b+1)})/2. \qquad (3.5)$$

Then

$$I_{rpb}=I_{temp} \text{ and } I_{rp(b+1)}=I_{temp}. \qquad (3.6)$$

The algorithm is described as below:

Input : I_{rpq}, r=1,...,M, p=1,…,n, q=1,..,K$_p$ (K$_p$ is varied)
Output :I_{rpq}, r=1,..,M, p=1,..,,N, q=1,..,K (K is fixed, and K < K$_p$)
Step1: For every r and p, and those unmerged image sequences, compute d$_i$, i=1,…,t-1, t is initialized as k$_p$
Step2: Find the minimum of d$_i$
Step3: $b = \arg\left(\min_i d_i\right), i = 1,...t-1$
Step4: Merge I_{rpb} and $I_{rp(b+1)}$
Step5: Tag I_{rpb} as merged, t=t-1
Step6: If t<k then go to step1

After Spatial-Temporal Image difference (STID) algorithm on all the image sequences, the resultant number of image sequences is fixed, say K. For every sentence r, the image sequences are listed as follows:

1st trained image sequences:I_{r11},I_{r12},-----,i_{r1k}
2nd trained image sequences:I_{r21},I_{r22},----I_{r2k}
………
Nth trained image sequences :I_{rn1},I_{rn2},----I_{rnk}

Average trained image sequences: $\hat{I}_{r1}, \hat{I}_{r2}, \cdots, \hat{I}_{rk}$

Where

$$\hat{I}_{rc} = \left(\sum_{y=1}^{N} I_{ryc}\right) \Big/ N, c = 1,...k \qquad (3.7)$$

Now the trained image sequences for sentences are as follows:

The 1st trained sentence: $\hat{I}_{11}, \hat{I}_{12}, \cdots, \hat{I}_{1k}$

The 2nd trained sentence: $\hat{I}_{21}, \hat{I}_{22}, \cdots, \hat{I}_{2k}$

....

The Mth trained sentence: $\hat{I}_{M1}, \hat{I}_{M2}, \cdots, \hat{I}_{Mk}$

Figure 3.6 shows the example trained image sequences after STID when K=10, where each column is the sequence number of image, and each row indicates a different sentence.

3.5 Recognition phase

Given a test image sequence $T_1, T_2,....,T_L$, we need to perform ***Spatial-Temporal Image difference (STID)*** algorithm before recognition phase. After STID, the test image sequences become $\hat{T}_1,\hat{T}_2,....,\hat{T}_k$. The Euclidean distance between the test image sequence and trained image sequences is used as a measurement of similarity. The Euclidean distance between the ith trained image sequences is defined as

$$E=\|\hat{T}_1-\hat{I}_{i1}\|+\|\hat{T}_2-\hat{I}_{i2}\|+....+\|\hat{T}_k-\hat{I}_{ik}\|, i=1,....,M, \quad (3.8)$$

where M is the number of sentences.

Then $\arg\left(\underset{i}{Min}(E_i)\right)$ is the index of the desired sentence to be recognized.

4. Three-Dimensional Head Model Reconstruction

Three-dimensional face model reconstruction from image sequences is a very challenging problem in computer vision and graphics. Our goal is to reconstruct a photo-realistic 3D face model in a short time by a regular digital camera and personal computer. After obtaining 3D head models of patients, we use them to establish a vivid communication system for hearing-impaired and speech-impaired patients. In the following subsections, techniques for establishing a 3D head model are first presented and the vivid communication system is finally explained and demonstrated.

4. Experimental Results

For the visual speech recognizer phase, the image sequences are taken from regular desktop CCD camera in the windows XP operating system. The image formats are RGB-based and are saved as a sequence of compressed JPEG images. The image sequences of multiple sentences are the training and testing data. Ten sentences were used as the trained sample data. Each sentence was spoken 10 times by the same person. The test data are also recorded by speaking 10 sentences and each sentence 10 times. Table 4.1 shows the confusion matrix of the test image sequences. From this table, the first sentence is recognized as the first sentence nine times and is recognized as the fourth sentence once. Sentence 3, 4, 7,and 9 are recognized perfectly. The recognition ratio for sentences 1,2,5,8, and 10 is 90%, and for sentence 6 is only 60%. The overall recognition ratio is 91% (for k=10). Table 4.2 shows the recognition rate versus k. Note that the recognition rate is proportional to the k value. Figure 4.1 shows the recognition rate versus the fixed number of frames after spatial-temporal algorithm.

Table 4.1: the confusion matrix of the test image sequence

	1	2	3	4	5	6	7	8	9	10
1	9				1					
2		9								
3			10							
4	1			10						
5		1			9					
6						6		1		
7							10			
8								9		
9						3			10	1
10						1				9

Table 4.2 : the recognition rate versus k

K	recognition rate
5	89%
10	91%
15	93%
20	94%
25	97%

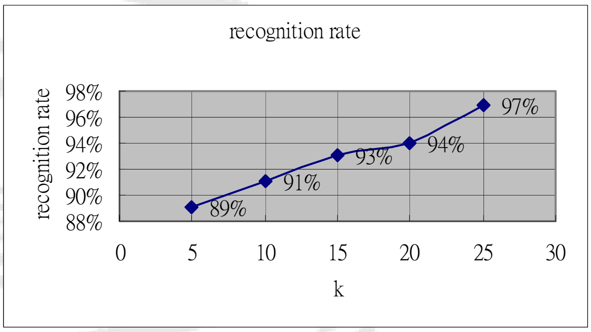

Figure 4.1: The recognition rate versus the fixed number of frames after spatial-temporal algorithm.

5. Conclusions and Future Works

After medical or surgical therapeutic processes, rehabilitation is the last resort to restore the capability of communication. For speech and hearing disorders, communication can usually be augmented by visual stimulation, for example, by signs or text. Owing to the development of the world-wide-web, not only can our traditional oro-aural communication mode be replaced by cyber-text and distant communication can also be substituted but the communication mode of the speech- or hearing-impaired people. For the proposed technique of data analysis, we can also build a visual speech recognition system for these patients, either for their post-operation communication or for their later speech rehabilitation period. Such system can also restore the patient's voice data before the operation and reconstruct their voice by the visual speech recognizing and visual analysis. The purpose of this research is to build a visual education system, which recognizes the sentence from the image sequences of hearing- or speech-impaired patients for communication and education.

The proposed system uses the spatial-temporal image difference technique to recognize limited sentences from the same speaker. The visual talking education tool is developed for speech- or hearing-impaired patients by showing the movement of lip to correct their pronunciation. In addition, we reconstruct 3D heads of patients in our visual communication system. The 3D animation system helps the hearing-impaired patients in communication.

Several works can be done in near future. For example, the proposed system should be capable of incorporating audio features for improving speech recognition system, especially in noisy background environment. The text-to-speech mechanism should be included in the near feature to enable speech- or hearing-impaired patients to communicate with normal people more naturally.

7. References

［1］ N.J. Cassisi, C Sapienza and B.P. Vinson, “ Malignant lesions of the larynx,” Organic voice disorders-Assessment and treatment, eds. W.S. Brown, B.P. Vinson and M.A.Crary, Singular Publishing Group, Inc, San Diego,1996, pp279-300.

［2］ C. Finiza, B. Bergman,” Health-related quality of life in patients with laryngeal cancer: A post-treatment comparison of different modes of communication”, The laryngoscope 2001, 111: 918-923.

［3］ Iain Mattews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, “Extraction of Visual Features for Lipreading”, IEEE Transactions on PAMI, Vol. 24, No. 2, Feb. 2002, pp. 198-213.

［4］ P.L. Silsbee and A.C. Bovik, “Computer lipreading for improved accuracy in automatic speech recognition,” IEEE Trans. Speech Audio Processing, Vol. 4, pp. 337-351, Sep. 1996

［5］ M. Gordan, C. Kotropoulos, and I. Pitas, “Pseudo automatic Lip Contour Detection Based on Edge Direction Patterns,” Image and Signal Processing and Analysis, ISPA 2001, pp. 138-143

［6］ M. Voget, “Fast Matching of a Dynamic Lip Model to Color Video Sequences under Regular Illumination Conditions,” Speech-Reading by Humans and Machines, 1996, pp. 399-407.

［7］ J. Luettin and N.A. Thacker, “Speech reading using Probabilistic Models,” Computer Vision and Image Understanding, 65, 2(1997), pp. 163-178.

［8］ C. Benoit, T. Guiard-Marigny, B. LeGoff and A. Adjoudani, “Which components of the face do humans and machines best speech read?” , Speech-Reading by Humans and Machines, 1996, pp. 351-372.

［9］ N.M. Brooke, “Talking heads and speech recognizers that can see: the computer processing of visual speech signals,”, Speech-Reading by Humans and Machines, 1996

［10］K. Yu, X. Tiang, and H. Bunke, ”Lipreading Using Signal Analysis over Time,” Signal Processing, 77(2), 1999, pp. 195-208.

［11］K. Yu, X. Jiang, and H. Bunke, “Sentence lipreading using Hidden Markov Model with integrated grammar,” Hidden Markov Models: Applications in Computer Vision (2001), World Scientific Publishing Company, pp. 161-176.

［12］Y. C. Lee, D. Terzopoulos, and K.Waters. “Realistic modeling for facial animation.” In Computer Graphics, Annual Conference Series, pages 55–62. SIGGRAPH, 1995.

［13］P. Marc, V.G. Luc, “Reading between the lines-a method for extracting dynamic 3D with texture.” Symposium on Virtual Reality Software and Technology’97, 95-102, 1997.

［14］ T. Akimoto, Y. Suenaga, and R. S. Wallace. “Automatic 3d facial models.” IEEE Computer Graphics and Applications.” 13(5):16–22, September 1993.

［15］H. H.S.Ip and L.Yin. “Constructing a 3d individualized head model from two orthogonal views.” The Visual Computer, (12):254–266, 1996

［16］ F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. “Synthesizing realistic facial expressions from photographs.” In Computer Graphics, Annual Conference Series, pages 75–84. Siggraph, July 1998.

［17］P. Fua, “Regularized Bundle-Adjustment to Model Heads from Image Sequences without Calibration Data” International Journal of Computer Vision, , 38(2). 153--171, 2000.

［18］ Zicheng Liu, Zhengyou Zhang, Chuck Jacobs, Michael Cohen, “Rapid Modeling of Animated Faces From Video.” Journal of Visualization and Computer Animation, Vol 12, No.4 (Sep. 2001), Page 227-240. Also in MSR-TR-2000-11

［19］N.P Chandrasiri, I. Barakonyi , T. Naemura , M. Ishizuak, and H. Harashima , “Communication over the internet using 3D agent with real-time facial expression analysis, synthesis and text to speech capabilities”, http://citeseer.nj.nec.com/chandrasiri02communication

［20］ E. Cosatto, et al., “Lifelike Talking Faces for Interactive Services”, Proceedings of the IEEE, Vol. 91, No. 9, September 2003, pp. 1406-1429.

［21］ E. Cosatto and P. Graf, ”Photo-Realistic Talking-Heads from Image Samples”, IEEE Trans. On Multimedia, Vol. 2, No. 3, September 2000, pp. 152-163.

［22］ E.Chung and Chris Bregler, “Performance Driven Facial Animation using Blendshape Interpolation”, Standford University Technical Report CS-TR-2002-02.

［23］Bergen et al., "Hierarchical model-based motion estimation" in Proceedings of Second European Conference on Computer Vision, Springer-Verlag, p. 237-252, 1992.

［24］H. Wang and S.F. Chang, “A Highly Efficient System for automatic face region detection in MPEG video”, IEEE Trans. On Circuits and Systems for Video Technology, Vol. 7, No. 4, Aug. 1997, pp. 615-628.

［25］Z. Zhang. “ A Flexible New Technique for Camera Calibration.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000.

［26］ Marquardt, D., “An Algorithm for Least Squares Estimation of Nonlinear Parameters,” SIAM J. Appl. Math. Vol. 11, pp. 431-441, 1963.

［27］ Reinhard Klette, Karsten Schluns and Andreas Koschan, “Computer Vision Three-Dimensional Data from Images”, ISBN 981-3083-71-9, Springer-Verlag Singapore Pte. Ltd. 1998.

［28］ R. C. Gonzalez and R. E. Woods, “Digital Image Processing (third edition), ” Reading, Massachusetts:

Addison-Wesley, 1992.

[29]D. Brown. “The bundle adjustment - progress and prospect”, XIII Congress of the ISPRS, Helsinki, 1976.

[30]R. Hathaway, “Another interpretation of the EM algorithm for mixture distributions”, Statistics and Probability Letters, 4:53-56, 1986.

[31]T. Fujii, T. Sato, K. Yoshino, K. Umatani, S. Ito and M. Hashimoto, “ Voice rehabilitation with esophageal speech in the laryngectomized”, Nippon Jibiinkoka Gakkai Kaiho, 1993, 96(7): 1086-93.

[32]J. Depondt, P. Gehanno, “ Laryngectomized patients’ education and follow-up”, Patient education and counseling, 1995, 26: 33-36.