Automatic Construction of Virtual Talking Faces and Applications

Cheng-Jyun Lai (賴成駿) and Wen-Hsiang Tsai (蔡文祥)
Department of Computer and Information Science,
National Chiao Tung University, Hsinchu, Taiwan, R. O. C.
Tel: 886-3-5712121 Ext. 56650, e-mail: gis91522@cis.nctu.edu.tw

Abstract

A system for automatic creation of virtual talking faces is proposed. The system is based on the use of 2D facial images and includes three major processes: video recording, feature learning, and animation generation. In the video recording process, a transcript containing all classes of Mandarin syllables is proposed. In the automatic feature learning process, a sentence segmentation algorithm is proposed to help the learning of audio features. Base image sequences that can exhibit natural head shaking actions are generated automatically. An image matching method is proposed to learn facial features with sub-pixel precision even on rotated faces. In the animation generation process, a method to conduct optimal superimposition of a mouth image onto a base image is proposed. To create more natural virtual faces, behaviors of real talking persons and singing persons are simulated. Good experimental results show the feasibility of the proposed methods.

1. Introduction

Although multimedia technologies for various applications have improved a lot, many people still feel unsatisfied and want their computers to have more human natures. One research topic, called virtual talking faces, concentrates on construction of simulated human faces with speaking capabilities on computer screens to help people interact with computers. Since people are used to interacting with others with their faces being seen, this research topic is of great use. Virtual talking faces can be applied to many areas, like virtual reporters, virtual receptionists, virtual teachers, virtual travel agents, etc.

In this study, we want to design an effective system for automatic creation of virtual talking faces for the Mandarin speech with realistic human appearances and fluent speaking abilities. The system, after learning a user's facial features, allows the user to input his/her speech, and will create accordingly an animated virtual face with moving lips uttering the input speech synchronously.

In Ezzat, Geiger, and Poggio [1], images are processed to synthesize new and previously unseen mouth configurations with the help of a multidimensional morphable model. Trajectories corresponding to desired utterances are synthesized. Then these mouth configurations are pasted onto background images to synthesize animations. In Cossato and Graf [2], trajectories of lips of recorded videos are analyzed to select best mouth images for the utterances. In Lin and Tsai [3], animations are created by rearrangements of recorded image frames. Image sequences of syllables are stretched to fit in the final animations. In King and Parent [4], differences between speeches and songs are analyzed, and the motion between the visemes of song singing are emphasized in the study.

2. System Overview

The proposed system consists of three main processes: video recording, feature learning, and animation generation. First, a model is asked to read aloud a pre-designed transcript with all Mandarin syllables on it, and the process is recorded as a video. Secondly, the video is analyzed to extract necessary feature information automatically. At last, the feature information and the speech data together are used to generate the final animation. The proposed methods make efforts in simplifying the video recording process, automating the feature learning process, and enhancing the qualities of the generated animation. A brief flowchart of the proposed system is illustrated in Fig. 1.

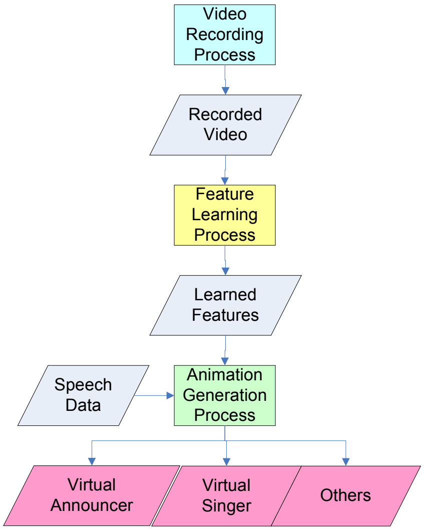


Fig. 1 Flowchart of proposed system.

3. Video Recording Process

The video recording process is designed to be as simple as possible in this study, so that a person (called a model hereafter) involved in the process will not feel confused or impatient. A scene of the environment setup for this study is described as follows: The model is seated in front of a camera. A pre-designed transcript is shown on a screen right behind the camera, and its position is adjusted in such a way that the model can read the transcript without obstacles. The entire process takes only about two minutes, which is quite short.

3.1. Transcript Reading

In this study, we make efforts to create virtual faces that are capable of speaking Chinese words. In [6], Lin and Tsai classified the 411 kinds of Mandarin syllables into 115 classes according to mouth shape similarities. However, speaking these syllables one by one singly is a boring work for the model. Therefore, we propose a transcript that contains the 115 classes of syllables by 17 sentences, as shown in Table 1. These sentences are designed to be meaningful and short so that the model can speak them easily. Efforts are also made to minimize repetitions of the syllables in this transcript.

Table 1 Proposed transcript that contains all kinds of Mandarin syllables.

Sentence	Used syllable classes
好朋友自遠方來	35, 63, 84, 2, 108, 51, 23
熟能生巧	39, 62, 59, 81
細水長流	66, 103, 46, 86
竊賊們否認行兇	76, 27, 57, 44, 53, 97, 115
歐洲平快車出發了	38, 39, 99, 102, 15, 69, 9, 18
難測風雲禍福	49, 16, 64, 109, 101, 71
春花三月分飛	105, 100, 47, 107, 58, 31
秋香百里撲鼻	85, 89, 24, 67, 70, 68
開滿森林更漂亮	22, 50, 54, 94, 61, 83, 90
刁傲丫頭惡劣荒謬	82, 32, 72, 42, 14, 77, 104, 87
民宅內一片黑暗	95, 20, 29, 65, 91, 28, 45
老翁和阿婆喝茶	36, 106, 48, 3, 12, 17, 4
別在崖下紮營唷	78, 21, 79, 73, 5, 96, 111
墾丁野馬愛吃嫩草	55, 98, 75, 8, 19, 1, 56, 34
信用卡被某人勾走	93, 115, 6, 30, 43, 53, 41, 40
誰要找陰陽佛塔	26, 80, 33, 92, 88, 13, 7
貓兒曾去報恩喔	37, 110, 60, 69, 37, 52, 10

3.2. Recording Process

Firstly, the model should keep his/her head facing straightly to the camera. Secondly, after the recording begins, the system operator should instruct the model to shake his/her head slightly for a predefined period of time while keeping silent. The recorded video of this time period is used as an assist for learning of audio features and construction of base image sequences. Thirdly, the model is instructed to read aloud the sentences on the transcript one after another, each followed by a predefined period of silent pause. These pauses are used to help learn audio features. An example of diagrams of recorded video contents and corresponding taken actions is shown in Fig 2.

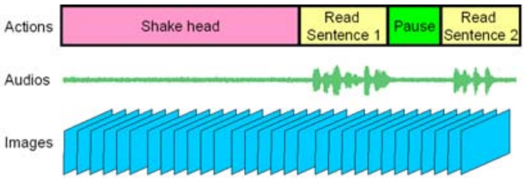


Fig. 2 An example of audios and images of recorded video, and corresponding actions.

4. Automatic Feature Learning Process

Features required for creation of virtual talking faces can be classified into three types: audio features, base image sequences, and facial features. Audio features are the timing information of the spoken syllables in the recorded video and are used to help synchronize the audios and the images. Base image sequences are the sequences of the background facial images onto which changeable facial parts such as mouths can be pasted to form faces that speak different words. Base image sequences also control the ways of head shaking. Facial features are special parts of faces, which can be used as natural marks. Since these learning processes require the processing of a lot of audio data and image frames, manual processes are not acceptable. Several methods for learning these features automatically are proposed in this study.

4.1. Learning of Audio Features

Since the pre-designed transcript is composed of 17 sentences designed in this study, the speech of every sentence must be learned before learning of the syllables. It is possible to learn the timing information of the syllables directly from the speech of the entire transcript without segmentation of the sentences. However, the work will take much more time while the length of the input audio increases.

In order to segment speeches of sentences automatically, "silence" features are used. The idea is to learn the feature of "silence" and decide the positions of silences, and then segment the sentences using the intermediate silence audio parts. Silence here means the recorded audio parts in which the model does not speak. Since the volume of these parts usually is not zero due to the environment noise, they cannot be detected by simply searching zero-volume zones. Instead, the maximum volume appeared in the period of head shaking is used as a threshold value. Then, the silent parts are found in this study by searching for the ones whose volumes are always smaller than the threshold value. Short pauses between syllables in a sentence should not be viewed as silences, so the lengths of the audio parts are put into consideration, too. The entire process of automatic sentence segmentation is as follows.

Step 1: Find the maximum volume V appearing in the audio parts within the head-shaking period D_{shake}.

Step 2: Find a continuous audio part $A_{silence}$

whose volume is always smaller than V and lasts longer than a predefined pausing period between sentences.

Step 3: Repeat Step 2 until all silent parts are collected.

Step 4: Find a continuous audio part $A_{sentence}$ that are not occupied by any $A_{silence}$.

Step 5: Repeat Step 4 until all sound parts are collected.

Step 6: Break $A_{transcript}$ into audio parts of sentences.

Fig. 3 shows a result of sentence segmentation, and the three sentences in the audio are found successfully. After the segmentation of sentences is done, the timing information of each syllable in a sentence can be learned by speech recognition techniques.

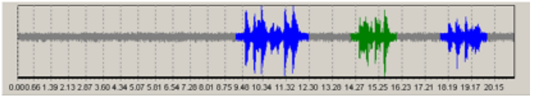


Fig. 3 A result of sentence segmentation.

4.2. Learning of Base Image Sequences

Base images provide places for variable facial features to be pasted on. For example, the mouths are pasted onto the base images in this study. The eyebrows and eyes on the base images are retained to produce animations with more natural eye-blinking actions. By inserting several images of a shaking head into the base image sequence, the produced animation can exhibit a speaking person with his/her head shaking naturally.

To produce a base image sequence, the initial silence period in the video recording process is utilized. The model is asked to shake his/her head during this period to simulate natural head shaking while speaking, and the image frames recorded during this period are used as base images. Since this period is short, it releases the model from fixing his/her eyes on the transcript during the entire recording process, which is a very tiring job.

An algorithm of proposed base image sequence generation is described as follows. It is noted that desired animations might require more image frames than the number of total base images, and so images in a sequence might be used repeatedly. One way to solve this problem is to reverse the traverse direction when reaching the first or the last base image. Fig. 4 illustrates this situation.

Step 1: Randomly select an initial frame $I_{initial}$ in the set of base images I.

Step 2: Randomly select an initial direction, either forward or backward.

Step 3: Add the current frame to the output sequence B.

Step 4: Stop learning if the number of frames in B are enough.

Step 5: Reverse the direction of traverse if the current frame is the first frame or the last in I.

Step 6: Advance to the next frame along the selected direction.

Step 7: Repeat Steps 3 through 6.

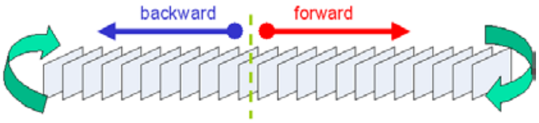


Fig. 4 A diagram of base image sequence generation.

4.3. Learning of Facial Features

To create an animation of a speaking person, the visemes, namely, the mouth images, should be pasted onto correct positions of faces; otherwise, the generated animation will look strange. In order to decide correct positions, a method based on image matching is proposed.

Firstly, a face detection technique using a knowledge-based approach is used in this study to learn the positions of facial features for the first frame. Common knowledge of facial features is used to detect their positions; however, sometimes errors occur because of some uncontrollable factors like lighting and shadows. Therefore, some manual modifications are allowed in our approach. Spatial relations between these features, which keep invariable for an identical face, are noted. Then, the nose, which is called the *base region* in this study, is detected by image-matching techniques. This base region remains as an invariant shape while a face is speaking. Finally, the positions of other facial features can be calculated according to these spatial relations. The process is illustrated in Fig. 5.

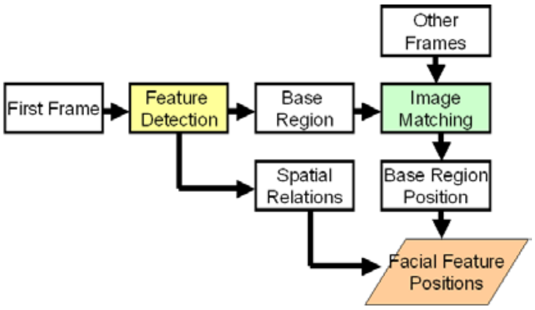


Fig. 5 Diagram of facial feature learning process.

For higher-precision animations, the positions of facial features should be learned with sub-pixel precision, since a face does not move one pixel each time; instead, more smoothly. That means, a pattern image needs to be shifted by a distance smaller than a pixel during image matching for facial feature learning. To accomplish the job of shifting, continuous properties of the facial images are utilized. Fig. 6 shows that after an image is acquired from a sensor, the coordinate values and amplitude values are sampled and quantized into digital forms, respectively. It is reasonable to "guess" that the amplitude values directed by the pink arrows shown in Fig 6 at the positions "between pixels" approximate the values of adjacent pixels since a face image is continuous. In this study, the technique of bilinear interpolation is used to generate these "sub-pixel values."

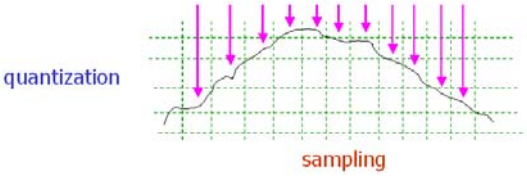


Fig. 6 Diagram of image quantization and sampling.

A method that performs image matching with sub-pixel precision using above-mentioned ideas is described as follows:

Step 1: Enlarge the pattern image $I_{pattern}$ r times using bilinear interpolation to get a new image $I_{patternL}$ where r is a predefined ratio.

Step 2: Enlarge the base image I_{base} r times using bilinear interpolation to get a new image I_{baseL}.

Step 3: Use the image matching technique to find the position $P'(x', y')$of $I_{patternL}$ in I_{baseL}.

Step 4: Divide x' by r to get x.

Step 5: Divide y' by r to get y.

For fixed faces, the above-mentioned method is suitable. However, for rotated faces, the base regions are quite different from the base region of the first frame and cause incorrect results of image matching. Another problem arises when the rotated angle of a rotated face is not known even if the position of the base region is detected correctly. For a straight face like Fig.

4.9(a), the positions of the facial features can be calculated correctly. However, for a rotated face like Fig. 4.9(c), the positions of facial features cannot be calculated correctly only with the help of the spatial relations even if the base region position is right.

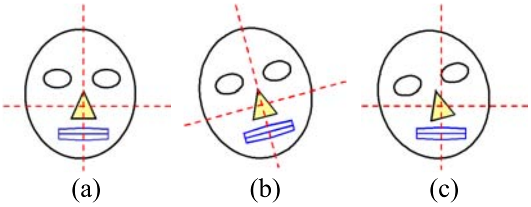


Fig. 7 Problem of image rotation.

To find the rotation angle of a rotated face, the base region image is rotated first to generate several rotated versions. All of these rotated images are used as input pattern images. And the image matching technique is applied. Finally, the best position and the corresponding rotation angle of the base region are obtained as the learning result.

5. Animation Generation Process

The proposed animation generation process goes as follows. First, a person is asked to read a transcript, and the speech is recorded. A process of syllable alignment is then applied to extract the timing information of the syllables in the speech. Proper image frames that are synchronized with the speech can be generated. Finally, animations are generated by composition of the images frames and the speech data. In the following sections, we described the efforts that are put in this study on the process of frame generation to create smoother animations.

5.1. Frame Generation of Speaking Case by Interpolation

To create synchronized animations, the number of frames for every syllable in an input audio is determined first; however, it may not equal the one of an identical syllable in the viseme database due to different speaking speeds and must be altered. To solve this problem, the technique of interpolation is used. The original frames are divided into N parts, where N is the number of desired frames. And then, a frame of each part is selected to represent the part. Finally, the content of the desired frames are replaced with the content of the representative frames one by one correspondingly.

It is also noticed that when a person speaks faster, his/her mouth shape changes more violently, and vice versa. To simulate this phenomenon, a process of frame generation for this case, called the speaking case, is designed and illustrated in Fig. 8. First, the mouth images of the syllables are determined by frame interpolation. Second, the visemes of the pauses between the syllables are decided. When the duration of a pause is long, it is considered that the person will close his/her mouth; otherwise, keep his/her mouth open and unchanged just like the situation that the pause does not exist. Third, the visemes of the first and the last pauses are restricted to be closed mouths.

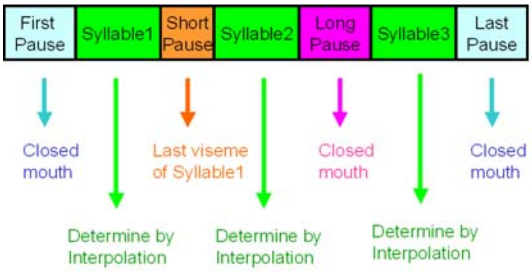


Fig. 8 Frame generation of the speaking case.

5.2 Frame Generation of Singing Case by Interpolation

Due to certain properties of songs, a singing person often has to utter syllables for longer time durations, especially when he/she is singing a slow song. The frame interpolation technique is not suitable for such a singing case because it will make a mouth shape change in slow motion, which is not natural. In solving this problem, several useful facts were found in this study. The first is that a mouth always keeps open while singing songs even during long pauses. The second is that after the sound of a syllable is uttered, the mouth will hold its shape unchanged and continue uttering the sound. Before the mouth holds its shape, we regard the person to be in a "mouth-opening" phase, and afterward, in a "mouth-holding" phase. Fig. 9 shows a diagram of these two phases.

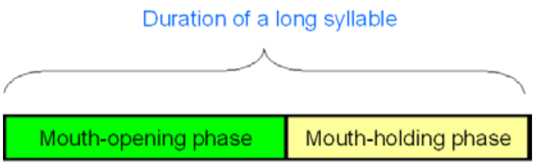


Fig. 9 Phases of a long syllable being sung.

The third fact is that the duration of the mouth-opening phase is related to the total duration of a syllable. When the duration of a syllable is longer, the duration of the mouth-opening phase is longer. Fig. 10 shows some experimental results proving this fact. The durations of the mouth-opening phase for different numbers of beats were observed. It can be seen indeed that when the duration of a syllable is longer, the duration of the mouth-opening phase becomes longer, too.

Therefore, the process of frame generation of the speaking case needs to be modified to fit these observed facts concerning singing songs. The first modification is that the mouth should not be closed during pauses. The second modification is about the two phases of singing. Suppose that the duration of a syllable S in the database is D_d, and that in an input audio of singing is D_a. When D_a is larger than D_d, the mouth should utter the sound in a duration of a certain value $D_{opening}$, and then keep its shape unchanged for a duration of $D_a - D_{opening}$. Here, the value of $D_{opening}$ is defined as

$$D_{opening} = D_d + \ D_a/\mathrm{D}_d$$

according to our experimental experience. And Fig. 11 shows the modified frame generation process of the singing case.

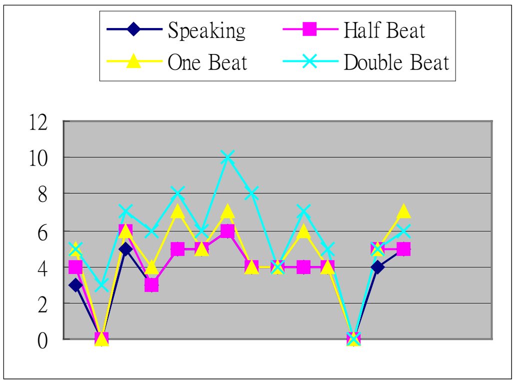


Fig. 10 Relations between singing speed (expressed by four beat numbers) and the number of frames of syllables for mouth opening (specified by the vertical axis). The horizontal axis indicates a sequence of uttered characters in a song.

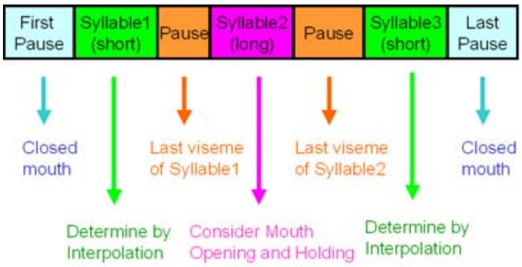


Fig. 11 Frame generation of the singing case.

5.2. Integration of Mouth Images and Base Images

After the frame generation process determines the visemes of an animation, the mouth image of a viseme can be integrated into a base image using the alpha-blending technique. Here a mouth image represents a region on a facial image that contains lips and a jaw. However, the determination of the region is not easy. In [3], Lin and Tsai used a fixed rectangle surrounding the mouth to represent the region. However, the determination of the size of the rectangle is not an easy job because it affects the integrated results severely. An excessively wide rectangle that overlaps the background such as walls may cause the background to be "integrated" into the face. An improperly small rectangle may cause the jaw to "drop down" while opening the mouth because only part of the jaw is moving.

A method is proposed in this study to determine the region of a mouth image automatically. Suppose that there are two images I_1 and I_2, and the mouth region of I_1 is to be integrated into I_2. A mouth region is a fixed rectangle surrounding the mouth. Assume that the mouth regions of I_1 and I_2 are M_1 and M_2, respectively. Then the proposed method goes as follows. First, the skin color is obtained by averaging the colors in the base region of I_1. Second, skin regions of M_1 and M_2 are determined, and denoted as S_1 and S_2, respectively. Finally, the intersection region $S_{\text{intersect}}$ of S_1 and S_2 is found and a region growing method is utilized to discard noise. $S_{\text{intersect}}$ is used as the mouth region. Alternatively, a trapezoid inside $S_{\text{intersect}}$ can also be used as another choice of the mouth region. Fig. 12 shows examples of these two kinds of mouth regions.

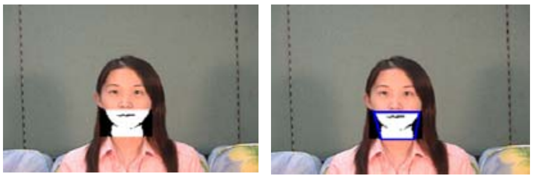


Fig. 12 The found mouth regions.

6. Experimental results

In this study, a system for automatic feature learning and animation creation has been constructed. Some experimental results of the proposed methods are shown here. For the learning of audio features, Fig. 13 shows the waveform of an audio, and blue and green parts represent odd and even sentences successfully found by the sentence segmentation algorithm.

For the learning of facial features, Fig. 14(a) shows an initial setup of the base region and the mouth region, which are represented by blue and yellow rectangles, respectively. The base region was determined by a knowledge-based face detection technique, and the mouth region was set to be a rectangle adjacent to the base region in this study. Fig. 14(b) shows a rotated face, and the proposed methods detected the position and rotation angle of the base region with sub-pixel precision successfully. Fig. 15(a) illustrates the influence of rotated faces. A mouth of a rotated face was integrated into a straight base face directly, which led to a bad integration result. In Fig. 15(b), the mouth was rotated in accordance with the base face first before the integration, and that produced a better result.

For the animation generation process, Fig. 16 shows some frames of a created animation in our experiment, and the face in the frames was speaking a Chinese sentence "熟能生巧."

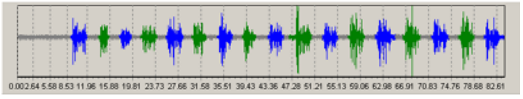
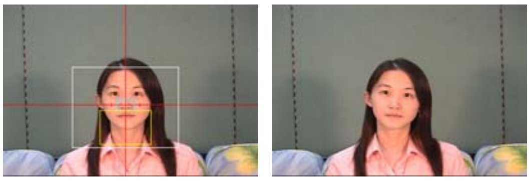
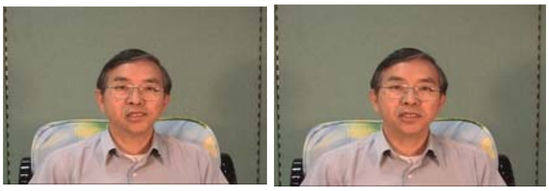


Fig. 13 An example of entire audio data of a transcript. The duration of head shaking is 5 seconds, and the duration of pausing between sentences is 1 second.



(a) (b)

Fig. 14 Illustration of face rotation. (a) A setup of the base region and mouth region. (b) A rotated face. The position of the base region is (348.0, 279.5), and the rotation angle is 6 degrees clockwise.



(a) (b)

Fig. 15 Illustration of necessity of face rotation before mouth integration. (a) A bad integration result. (b) A better result.

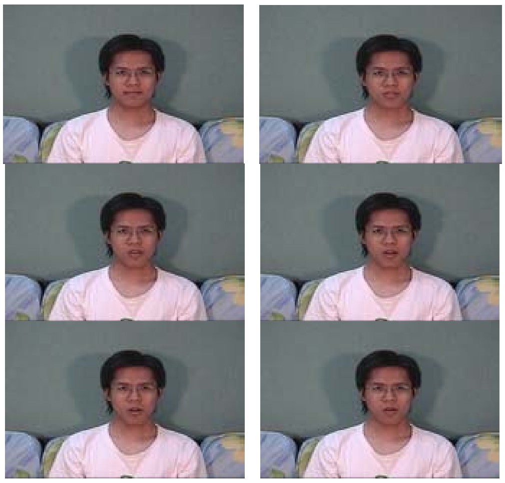


Fig. 16 An example of experimental results.

7. Conclusions

In this study, a system for creating virtual talking faces has been implemented. The system is based on the use of 2D facial images. Proper methods have proposed to automate the learning process and improve the quality of generated animations.

The proposed video recording process is short, easy, and not annoying. A transcript that contains all classes of Mandarin syllables has been proposed to help ease the model. The model is allowed to shake his/her head slightly, yielding more natural animation results. In the feature learning process, audio features, facial features, and base image sequences are all learned automatically. The learning process of facial features even can work well on rotated faces. In the animation generation process, the behaviors of talking persons and singing ones were also analyzed, and a method of frame generation that is proper to create both talking and singing faces was proposed. An interesting topic for future works is real-time animations of virtual talking heads with on-line speech inputs.

REFERENCES

[1] T. Ezzat, G. Geiger, and T. Poggio, “Trainable Videorealistic Speech Animation,” *Proceedings of SIGGRAPH*, San Antonio, Texas, USA, July 21-26, 2002.

[2] E. Cosatto and H. P. Graf, “Photo-Realistic Talking-Heads from Image Samples,” *IEEE Transactions on Multimedia*, Vol. 2, No. 3, Sept. 2000

[3] Y. C. Lin and W. H. Tsai, “A Study on Virtual Talking Head Animation by 2D Image Analysis and Voice Synchronization Techniques,” *M. S. Thesis*, Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, Republic of China, June 2002.

[4] S. A. King and R. E. Parent, “Lip Synchronization for Song,” *Proceedings of 15th International Conference On Computer Animation (Computer Animation 2002),* Geneva, Switzerland, June 19-21, 2002.

[5] W. S. Lee and N. M. Thalmann, “Generating a Population of Animated faces from Pictures,” *Proceedings of IEEE International Workshop on Modelling People*, Corfu, Greece, Sept. 20-20, 1999, pp. 62-62.