國立交通大學

資訊科學系

碩士論文

自動化建構虛擬說話人臉與其相關應用之研究

A Study on Automatic Construction of Virtual Talking Faces and Applications

研究生：賴成駿

指導教授：蔡文祥　教授

中華民國九十三年六月

自動化建構虛擬說話人臉與其相關應用之研究
A Study on Automatic Construction of Virtual Talking Faces and Applications

研究生：賴成駿　　Student：Cheng-Jyun Lai

指導教授：蔡文祥　　Advisor：Wen-Hsiang Tsai

國立交通大學
資訊科學研究所
碩士論文

A Thesis
Submitted to Institute of Computer and Information Science
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer and Information Science

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

自動化建構虛擬說話人臉與其相關應用之研究

研究生： 賴成駿　　　　　　　　　指導教授： 蔡文祥 博士

國立交通大學資訊科學研究所

摘 要

本論文提出了一套自動化建構虛擬說話人臉的系統。這個系統以二維臉部影像為基礎，包含了三個階段：錄影學習、特徵值學習與動畫製作。在錄影學習階段，我們提出了一個包含所有種類的中文注音的稿子，模特兒只要唸上面的句子就可以完成學習，而不用單獨唸每個音。在特徵值學習階段，語音特徵、臉部特徵與背景影像序列等資訊系統都會自動學習，並以自動斷句來輔助學習語音特徵。另本系統亦能產生自然搖頭效果的背景影像序列，基於影像比對方法學習臉部特徵的位置。在達到次像素精準度的同時，這個方法也可以適用在搖動的人臉上。在動畫製作階段，我們提出了幾個方法來增進動畫的精細度。首先提出了一個達成語音與影像同步的方法。為了建立更流暢的動畫，我們分析了相連音間所轉折畫格數目，也提出了一個自動決定嘴巴影像與背景影像最佳整合方式的方法。為了建立更真實的虛擬人臉，我們研究並模擬了真人說話和唱歌時的行為。最後我們實作出三種有趣的應用。良好的實驗結果證實本論文所提出方法之可行性。

A Study on Automatic Construction of Virtual Talking Faces and Applications

Student: Cheng-Jyun Lai Advisor: Dr. Wen-Hsiang Tsai

Institute of Computer and Information Science
National Chiao Tung University

ABSTRACT

In this study, 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 processes: video recording, feature learning, and animation generation. In the video recording process, a transcript containing all classes of Mandarin syllables is proposed, so that a model can read sentences on it instead of reading all the syllables separately. In the feature learning process, audio features, facial features, and base image sequences are all learned automatically. A sentence segmentation algorithm is proposed to help the learning of syllables. Base image sequences that can exhibit natural head shaking actions are generated. An image matching method is proposed to learn the positions of facial features in a face image with sub-pixel precision. The method also can be applied to shaking faces. In the animation generation process, several methods are proposed to improve the quality of animations. A method is proposed to synchronize a speech and image frames. To create smoother animations, the number of proper transition frames between successive visemes is analyzed. Also proposed is method to find the best way for integration of a mouth image and a base image. To create more natural virtual faces, a method is proposed to simulate the behaviors of real talking persons and singing persons. Three kinds of interesting applications are implemented. Good experimental results show the feasibility of the proposed methods.

ACKNOWLEDGEMENTS

The author is in hearty appreciation of the continuous guidance, discussions, support, and encouragement received from his advisor, Dr. Wen-Hsiang Tsai, not only in the development of this thesis, but also in every aspect of his personal growth.

Thanks are due to Mr. Chih-Hsuan Tzeng, Mr. Chang-Chou Lin, Mr. Chih-Jen Wu, Mr. Tsung-Yuan Liu, Ms. Yen-Lin Chen, Mr. Yen-Chung Chiu, Mr. Nan-Kun Lo, Mr. Wei-Liang Lin, Mr. Yi-Chieh Chen and Mr. Kuei-Li Huang for their valuable discussions, suggestions, and encouragement. Appreciation is also given to the colleagues of the Computer Vision Laboratory in the Department of Computer and Information Science at National Chiao Tung University for their suggestions and help during his thesis study.

Finally, the author also extends his profound thanks to his family for their lasting love, care, and encouragement. He dedicates this dissertation to his parents.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 Motivation

In recent years, communication technologies have been improved a lot, inspired by the exploding amount of communication activities on the Internet. Much more information can be transmitted through networks, and so promotes improvement of multimedia technologies. People are now used to watching high-quality films and playing computer games with pretty appearances. However, these technologies do not make computers friendlier. More and more people feel unsatisfied and want their computers to have more human natures.

Due to this expectation, more and more researchers start to invest their efforts in improvements of interfaces between humans and computers. One of the research topics, called virtual talking faces, concentrates on reconstructions of human faces on computer screens. With the technologies of virtual talking faces, people can watch, listen, or even speak to virtual humans on their computer screens. Since people are used to interacting with others with their faces being seen, this research topic is of great use.

The first famous character created by the technologies of virtual talking faces is Ananova, a television announcer. She is able to report news in fluent English, just like she were a real television announcer. Although her face is not real and pretty enough, she still attracts much attention on the Internet. Before the appearance of Ananova, news reports on the Internet were written in texts. It is unprecedented for people to listen and watch real-time news reporting by a virtual announcer on the Internet.

Encouraged by the success of Ananova, the technologies of virtual talking faces are being applied to more areas. For example, navigation software uses virtual characters to help illustrate tour maps and give traveling suggestions. Web masters use virtual characters to help illustrate their web sites, goods, and services. Corporations use virtual characters to help handle calls on telephone exchanges.

Besides improvements on interaction between computers and humans, virtual talking faces have other functions. One of them is to save required bandwidth while transmitting videos of talking persons by networks, for example, video conferencing. Since videos of virtual talking faces can be transmitted by networks with only a small data amount and then be reconstructed by receivers without the need of transmitting all image frames of animations, the requirement of the bandwidth is much lower than those required for transmissions of normal movies. Virtual faces are also used in Rickel and Marsella [1] as tutors.

In order to achieve the above-mentioned goals, virtual faces at least must have two properties, namely, realistic appearances and fluent speaking capabilities. To create virtual faces with realistic appearances, there are two main approaches. The first is to create virtual faces using 3D head models. One advantage of this approach is that created virtual faces have fewer limits on head movements. However, they bear the disadvantage of yielding rougher facial appearances, which is usually not acceptable. In the second approach, virtual faces are constructed using 2D image samples. Though producing better facial appearances, they have the disadvantage of imposing more limits on possible applications.

Speaking fluently is another important and essential property of a good virtual talking face. This property relies on the help of many technologies, such as speech recognition and synthesis. Language is another important issue. Virtual faces that are able to speak various kinds of languages are preferred, however, hard to implement.

In this study, we want to design an effective system for automatic creation of virtual talking faces for the Mandarin speech, which possesses the above-mentioned good properties. First, we want to establish a database of essential features. Users can then input their speech, and the system will create an animated virtual face with moving lips uttering the input Mandarin speech synchronously.

1.2 Survey of Related Studies

As mentioned above, there are two main approaches to creation of virtual faces. In the first, virtual faces are created by the use of 3D head models. For example, a generic 3D face is used in Zhang and Cohen [2]. Multiple image views of a human face are utilized to morph the generic face into specific face structures. In Goto, Kshirsagar, and Thalmann [3], two orthogonal pictures of a human are taken, and a generic 3D face model is then modified to fit the pictures.

In the second approach, virtual faces are created by the use of 2D image samples. In Ezzat, Geiger, and Poggio [4], 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 [5], trajectories of lips of recorded videos are analyzed to select best mouth images for the utterances. In Lin and Tsai [6], animations are created by rearrangements of recorded image frames. Image sequences of syllables are stretched to fit in the final animations.

In order to let virtual faces be able to sing songs, differences between speech and songs must be analyzed. In [7], King and Parent believed that the motion between the visemes become extremely important while singing songs. In this study, we also investigate the topic of virtual singers, but with emphasis on Mandarin song synchronization that is less studied in the past.

1.3 Overview of Proposed Methods

Overviews of the proposed methods are described in this section. First, some terms used in this study are defined in Section 1.3.1. And some assumptions made for this study are listed in Section 1.3.2. And at last, some brief descriptions of the proposed methods are described in Section 1.3.3.

1.3.1 Definitions of Terms

The definitions of some terms used in this study are listed as follows.

(1) *Animation:* An animation is a video created by a system as a final result. In the video, a realistic virtual face speaks some sentences or sings some songs.

(2) *Base Image Sequence*: A base image sequence is a sequence of images containing faces onto which some variable facial features may be pasted to form final animations.

(3) *Facial Feature:* A facial feature is a particular region on a face, which can be used as a mark, such as an eye, a nose, a mouth, etc.

(4) *Hidden Markov Model (HMM)*: The HMM is used to characterize the spectral properties of the frames of a speech pattern.

(5) *Mandarin Speech Database across Taiwan (MAT)*: The MAT is a database that collects voices through telephone networks from Mandarin speakers of different genders and ages in Taiwan.

(6) *Model*: A model is a person, either male or female, whose actions are recorded

in the learning stage. Faces of the model are used to create final animations.

(7) *Phoneme*: A phoneme is a basic enunciation of a language, like ㄅ, ㄊ, ㄨ, ㄤ in Mandarin.

(8) *Speech Analyzer*: A speech analyzer accepts a speech and a script as input, and utilizes speech recognition techniques to get the timing information of every syllable.

(9) *Syllable*: A syllable consists of one or more phonemes like ㄅㄨ, ㄉㄠ in Mandarin.

(10) *Transcript*: A transcript is a text file that contains the corresponding content of a speech.

(11) *Viseme*: A viseme is the visual counterpart of a syllable.

1.3.2 Assumptions

In the proposed system, video and audio data are acquired from a single camera. The properties of these data may be influenced by many factors. For example, noise in the audio might affect the result of syllable segmentation severely. And variations in lighting sometimes will cause complicated changes in the images.

In order to reduce the complexity of processing works in the proposed system, some assumptions are made in this study, which are described as follows.

(1) The environment is noiseless.

(2) The speech is spoken at a steady speed and in a loud voice.

(3) The lighting of the environment is constant.

(4) The face of the model always faces the camera and is located on the center of the recorded frame.

(5) The motions of the model are slight and slow.

1.3.3 Brief Descriptions of Proposed Methods

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 into a video. Secondly, the video is analyzed to extract necessary feature information. At last, the feature information and the speech data together are used to generate the final animation. The animation may have many forms, such as a virtual announcer, a virtual singer, a virtual teacher, etc. The proposed methods make efforts in simplifying the video recording process, automating the feature learning process, enhancing the qualities of the resulting animation generation process, and expanding possible applications.

A brief flowchart of the proposed system is illustrated in Fig. 1.1. The details of the principle behind the system and its configuration will be explained in Chapter 2.

1.4 Contributions

Some major contributions of this study are listed as follows.

(1) A complete system for creating virtual talking faces automatically is proposed.

(2) A transcript containing all Mandarin syllables is proposed.

(3) Some methods for gathering audio features automatically are proposed.

(4) Some methods for gathering facial features automatically are proposed.

(5) A method for gathering base image sequences is proposed.

(6) Several methods for improving the qualities of the final animations are proposed.

(7) Several new applications are proposed of the proposed system and implemented.

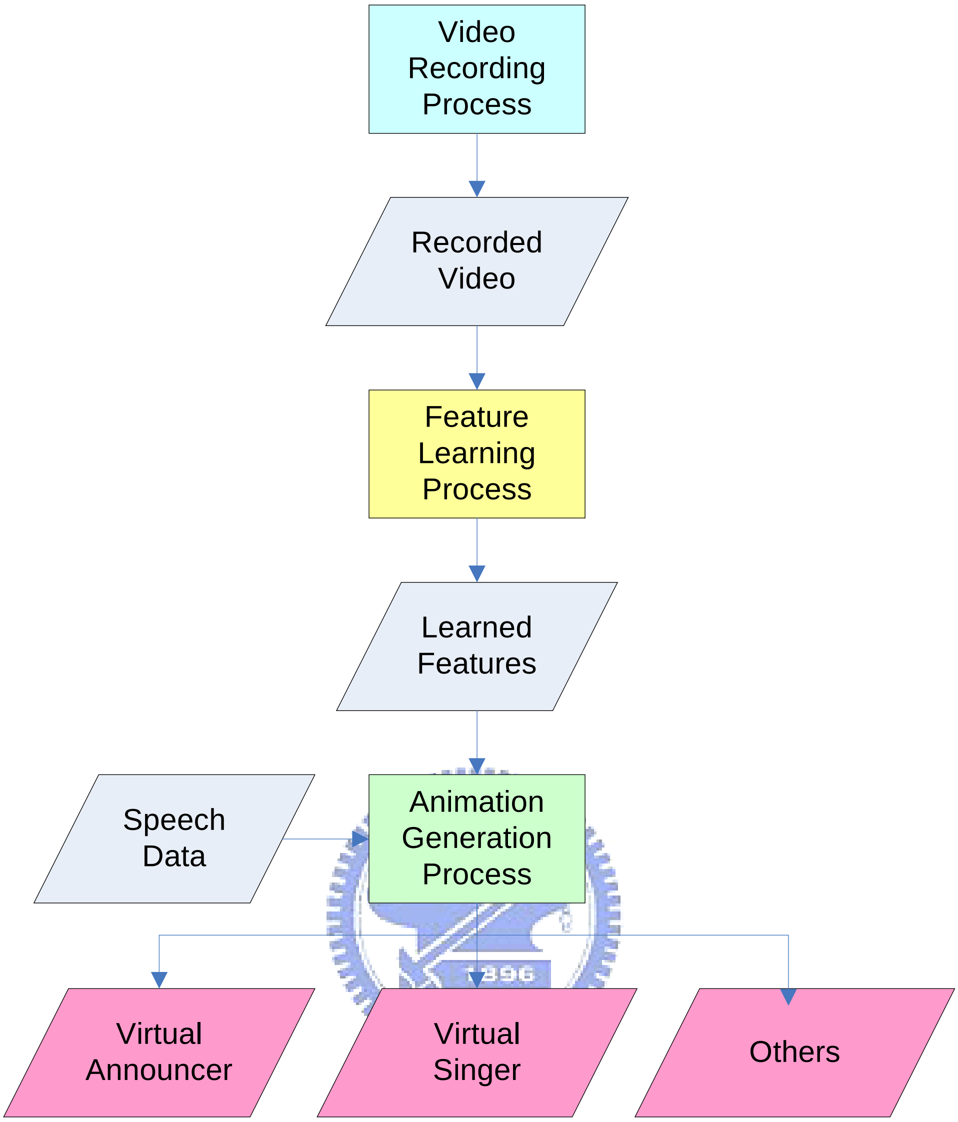


Fig. 1.1 Flowchart of proposed system.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 describes an overview of the proposed system and processes. Chapter 3 presents the proposed methods for learning audio features automatically. Chapter 4 presents the proposed methods to locate base regions automatically and precisely. Chapter 5 describes the proposed methods to locate other facial features automatically. Chapter 6 presents the proposed methods to generate smooth animations. In Chapter 7, some applications using the proposed methods are presented. Finally, conclusions and some suggestions for future research are included in Chapter 8. Experimental results and discussions are given in each chapter.

Chapter 2
System Overview

2.1 System Organization and Processes

As illustrated in Fig 1.1, the proposed system consists of three main processes: video recording, feature learning, and animation generation. In this section, relations between those processes are described.

For the video recording process, an output of a video containing a speaking model (a person) is desired. The process must be designed carefully, so that following processes can gather enough information from the video for creation of virtual talking faces. The process must also be simple and reasonable. Otherwise, the model may feel uncomfortable. Fig 2.1(a) illustrates the proposed video recording process.

For the feature learning process, an output of facial feature information is desired. The output of the video recording process is used as the only input. In order to reduce artificial interferences, several methods for automatic extraction of different kinds of features are proposed. Fig 2.1(b) illustrates the feature learning process.

For the animation creation process, outputs of animations containing virtual talking faces are desired. Sound tracks of speech are used as inputs, which control timing information of spoken syllables. Feature information from the feature learning process is also used as an input to help generate final animations. The generated animations must be smooth and realistic. Otherwise, viewers will not be satisfied. The animations may have different forms, such as virtual announcers, virtual singers, virtual teachers, etc. Fig 2.1(c) illustrates the animation creation process.

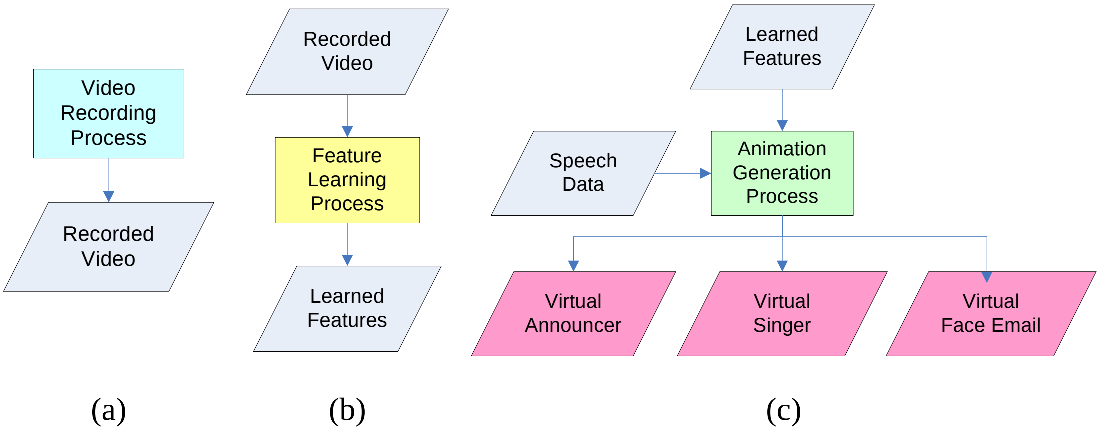


Fig. 2.1 Flowcharts of three main processes. (a) Video recording process. (b) Feature learning process. (c) Animation generation process.

In the following sections, contents of these processes will be described in detail.

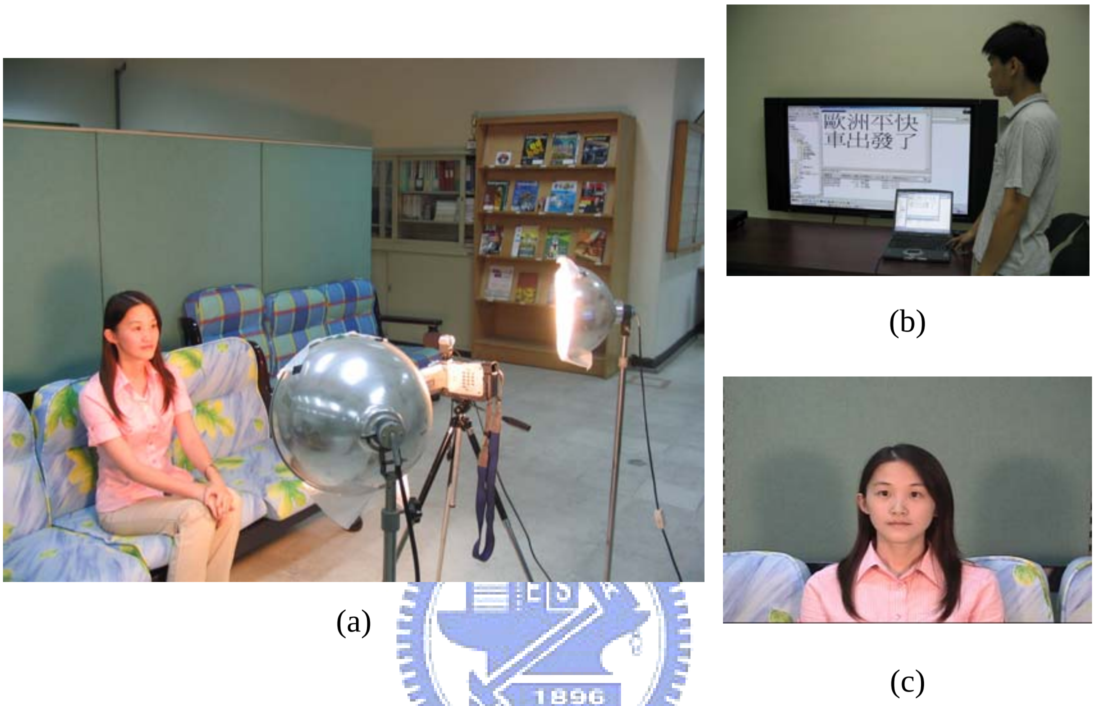
2.2 Video Recording Process

In the following sections, matters needing attention in the video recording process are described in detail. In Section 2.2.1, setups of recording environments are discussed. In Section 2.2.2, a transcript for learning of Mandarin syllables is proposed. In Section 2.2.3, a detailed recording process is proposed.

2.2.1 Environment Setup

The arrangement of the recording environment affects impressions of models severely. Since the models may not be familiar friends of system operators, the recording process should be as simple as possible, so that they will not feel confused or impatient. A too lengthy or complicated recording process is not acceptable.

A scene of our environment setup is shown in Fig 2.2. The proposed recording environment setup is rather simple. 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 so that the model can read the transcript without obstacles. As we can see in Chapter 8, the process takes about 2 minutes only, which is quite short.



(a)

(b)

(c)

Fig. 2.2 Scene of proposed environment setup. (a) The model. (b) The transcript. (c) The recorded scene.

In order to capture videos of better qualities, some extra devices are adopted in our environment setup. Introducing these devices doesn’t affect the simplicity of the process, but only add a little workload on system operators. For example, instead of normal webcams, a DigitalVideo device, which is capable of grabbing frames of 720×480 dimensions on a frame rate of 29.97/sec, is used to record videos. Two spotlights are also used. They not only brighten the model, but also reduce blinking effects of fluorescent lamps. Since final animations are generated from recorded frames, an environment with steadier lighting makes the resulting animations look smoother.

2.2.2 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. For virtual faces that are able to speak all Mandarin words, learning of these 115 classes of Mandarin syllables is necessary. However, speaking these syllables one by one singly is a somewhat boring work for models. Therefore, we propose a transcript that contains the 115 classes of syllables by 17 sentences, which is shown in Table 2.1. These sentences are designed to be meaningful and short, so models can speak them easily. Efforts are also made to minimize repetitions of syllables in this transcript.

Table 2.1 The proposed transcript that contains 115 classes of Mandarin syllables.

Number	Sentence	Used syllable classes
1	好朋友自遠方來	35、63、84、2、108、51、23
2	熟能生巧	39、62、59、81
3	細水長流	66、103、46、86
4	竊賊們否認行兇	76、27、57、44、53、97、115
5	歐洲平快車出發了	38、39、99、102、15、69、9、18
6	難測風雲禍福	49、16、64、109、101、71
7	春花三月分飛	105、100、47、107、58、31
8	秋香百里撲鼻	85、89、24、67、70、68
9	開滿森林更漂亮	22、50、54、94、61、83、90
10	刁傲丫頭惡劣荒謬	82、32、72、42、14、77、104、87
11	民宅內一片黑暗	95、20、29、65、91、28、45

Table 2.1 The proposed transcript that contains 115 classes of Mandarin syllables. (Continue)

Number	Sentence	Used syllable classes
12	老翁和阿婆喝茶	36、106、48、3、12、17、4
13	別在崖下紮營唷	78、21、79、73、5、96、111
14	墾丁野馬愛吃嫩草	55、98、75、8、19、1、56、34
15	信用卡被某人勾走	93、115、6、30、43、53、41、40
16	誰要找陰陽佛塔	26、80、33、92、88、13、7
17	貓兒曾去報恩喔	37、110、60、69、37、52、10

2.2.3 Recording Process

After the environment is set up and the transcript is prepared, the recording process can begin. For the convenience of feature learning, some extra works must be done in the recording process. However, these works add only a little workload to system operators and models, which are acceptable.

Firstly, the model should keep his/her head straight to the camera, and then the recording process can begin. Since the first frame of the recorded video is used as a reference frame in the feature learning process, a “straight” face with a normal expression is required. Otherwise, poorer information may be learned in the subsequent feature learning process.

Secondly, after the recording begins, the system operator should instruct the model to shake his/her head for a predefined period of time while keeping silent. The recorded video of this period of time is used as an assist for learning of audio features and base image sequences, which will be described in the following section.

Thirdly, the model is instructed to read aloud the sentences on the transcript one by one, each followed by a predefined period of silent pause. These pauses are used to help learn audio features. The model should read these sentences loudly, clearly, and slowly, so that syllables can be learned correctly.

A flowchart of the video recording process is illustrated in Fig 2.3. A diagram of the content of the recorded video and the corresponding actions taken is shown in Fig 2.4.

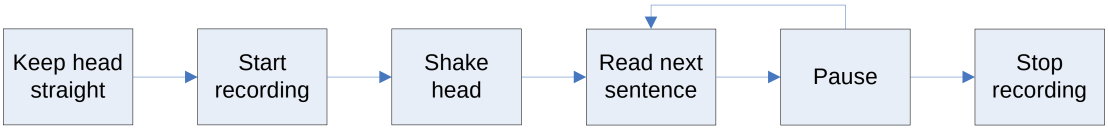


Fig. 2.3 A flowchart of the video recording process.

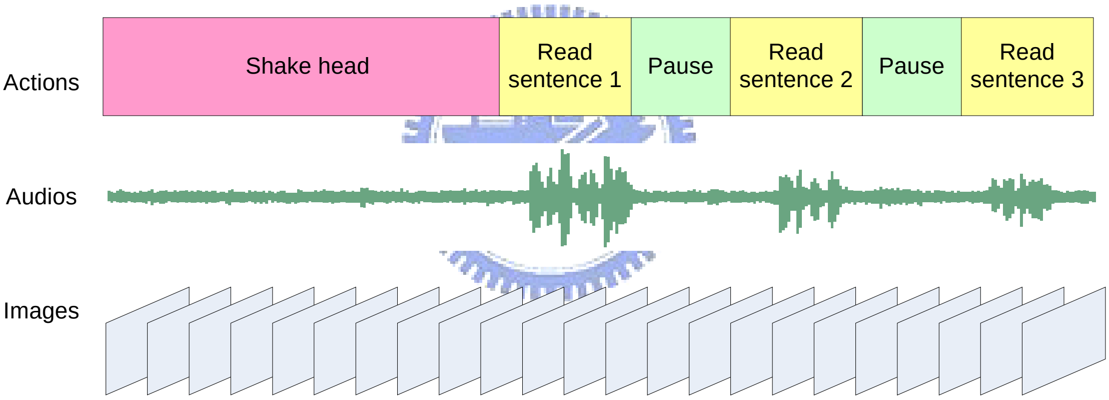


Fig. 2.4 A diagram showing the audios and images of the recorded video, and the corresponding actions taken.

2.3 Feature Learning Process

After the video recording process is done, features can be learned from the recorded video. Section 2.3.1 lists required features in the proposed system, and Section 2.3.2 illustrates the learning process for these different features.

2.3.1 Feature Classification

Features required for creation of virtual talking faces can be classified into four types: audio features, base image sequences, facial features, and base regions.

Audio features are timing information of spoken syllables in the recorded video. Timing information of the total speech, timing information of each sentence, and timing information of each syllable are examples of audio features. These features are used to help synchronize audios and images.

Base image sequences are sequences of facial images, which are used as background images. Mutable facial parts such as mouths can be pasted onto these images to form faces that speak different words. Base image sequences also control ways of shaking heads.

Facial features are special parts of faces, which can be used as natural marks. For example, noses, lips, and jaws are facial features adopted in this study.

Base regions are special facial features that can be used to orient faces. With the help of base regions, the positions and gradients of faces can be calculated. In this study, noses are adopted as the base regions to locate faces.

2.3.2 Learning Process

Fig 2.5 illustrates a flowchart of the feature learning process. First, the recorded video is split into audio data and image frames. With the help of the transcript, audio features can be learned from the audio data. Facial features can be learned directly from image frames. Learning of base image sequences requires the information of both audio data and image frames.

Since these learning processes require dealing with a lot of audio data and image frames, manual processes are not acceptable. Several methods for learning these features automatically are proposed and explained in Chapters 3 and 4.

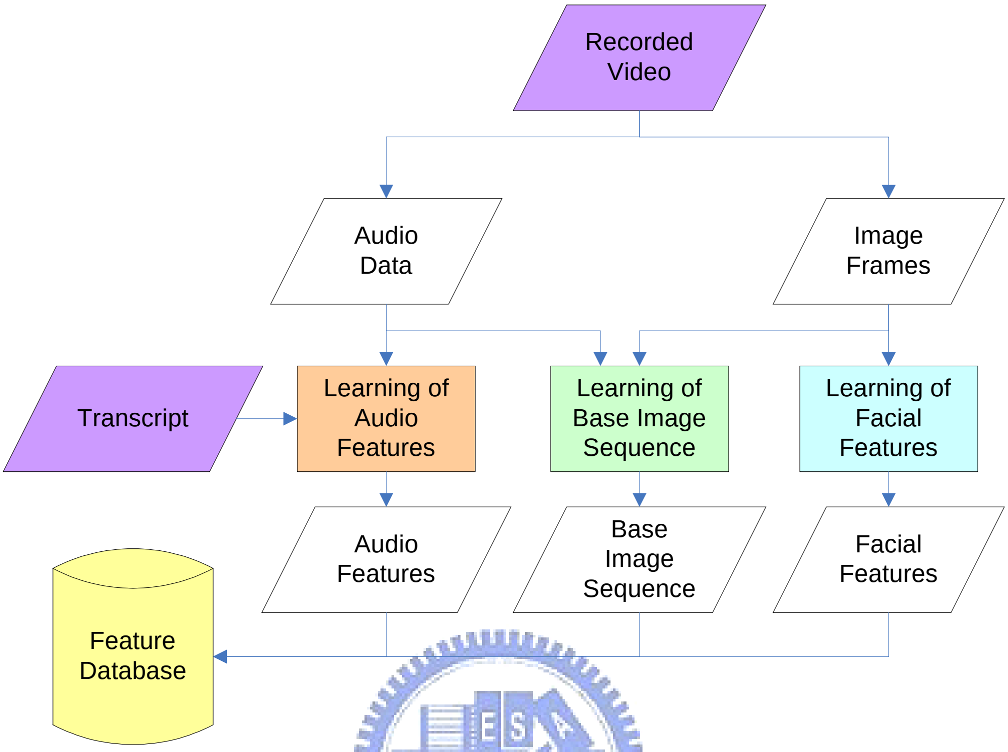


Fig. 2.5 A flowchart of the feature learning process.

2.4 Animation Generation Process

After features have been collected, animations of virtual faces can be created. In Section 2.4.1, some essential properties for animations are described. The animation generation process is illustrated in Section 2.4.2.

2.4.1 Properties of Animations

To create virtual faces that are capable of improving interfaces between humans and computers, created animations should possess several properties. Firstly, they should have realistic appearances. Interacting with faces that are not realistic is a very strange and unnatural thing. Secondly, lip movements should be smooth. Since people are familiar with watching others' lip movements while talking to each other, unnatural movements of lips will be discovered easily. Thirdly, fluent speaking abilities are required. And fourthly, speech and lip movements should be synchronized. Humans are conscious of asynchronous problems between speech and lip movements.

In this study, created animations by the proposed system possess all of these properties. Firstly, animations are generated from 2D image frames. With good techniques for integration of facial parts, animations look realistic and natural. Secondly, several methods are proposed to smooth the lip movements. Thirdly, original sound tracks of real people are adopted in final animations, which avoid the problem of unnatural voices. Fourthly, a method is proposed to synchronize speech and lip movements.

2.4.2 Animation Generation Process

The animation generation process requires two inputs: a transcript and its corresponding speech data. First, a person is asked to read a transcript, and the speech is recorded. The process of syllable alignment then extracts the timing information of the syllables in the speech. With the help of timing information of syllables and feature data, proper image frames that are synchronized with speech can be generated. Finally, animations are generated by composition of images frames and speech data.

Fig 2.6 illustrates a flowchart of the animation generation process.

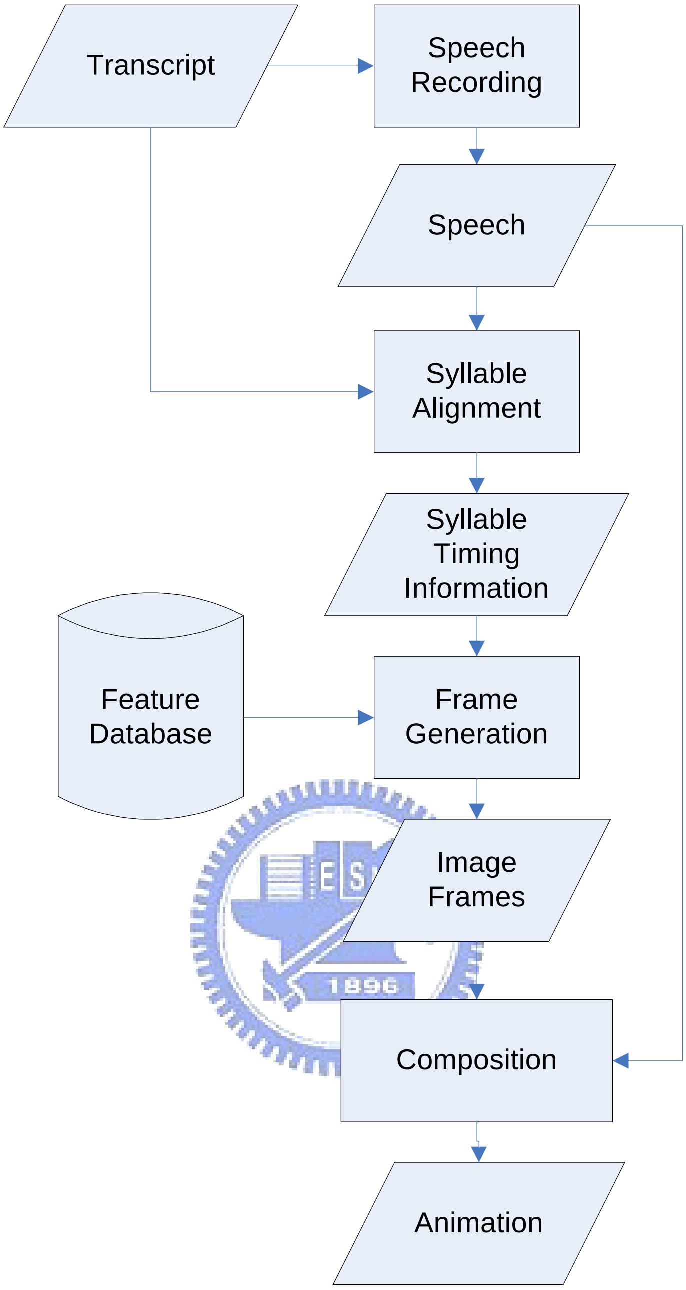


Fig. 2.6 A flowchart of the animation generation process.

Chapter 3
Automatic Learning of Audio Features and Base Image Sequences

3.1 Learning of Audio Features

In Section 2.3 in the last chapter, four types of features that must be learned in the feature learning process have been described. In the following sections, concentration is put on the learning of audio features. In Section 3.1.1, the audio features used in this study are described in detail. In Section 3.1.2, a method for segmentation of sentences is proposed. And in Section 3.1.3, the process of syllable alignment is reviewed.

3.1.1 Descriptions of Audio Features

In the video recording process, a video of the face of a model containing a speech of a pre-designed transcript is recorded. The speech includes the timing information, namely the duration, of every syllable that must be learned. Without the information, the work of assigning syllable labels to image frames cannot be done, and this makes it impossible to know which syllable an image frame belongs to.

Since the pre-designed transcript is composed of seventeen sentences designed in this study, the speech of every sentence must be learned first, before the 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. By segmenting sentences in advance, shorter audio parts are used in the learning process, which accelerates the processing speed.

Audio features mentioned above are listed in Table 3.1.

Table 3.1 Descriptions of audio features.

Feature	Description	Example
Speech of Transcript	A speech that contains the audio data of the entire transcript including seventeen sentences.	好朋友自遠方來。熟能生巧…貓兒曾去報恩喔。
Speech of Sentence	A speech that contains the audio data of a single sentence including several syllables.	好朋友自遠方來。
Speech of Syllable	A speech that contains the audio data of a single syllable.	ㄏㄠ、ㄆㄥ、ㄧㄡ、ㄗ、ㄩㄢ、ㄈㄤ、ㄌㄞ

3.1.2 Segmentation of Sentences by Silence Features

In the preceding section, the reason for sentence segmentation was explained. In the following sections, a method for sentence segmentation is proposed. A new kind of audio feature called “silence feature” used in the proposed method is introduced first.

3.1.2.1 Idea of Segmentation

In order to segment speeches of sentences automatically, the video recording process is designed to let the model keep silent for predefined periods of time in two situations as defined in the following:

(1) *Initial silence*: A period of time when the model keeps silent while shaking

his/her head, such as the red part in Fig 3.1.

(2) *Intermediate silence*: A period of time when the model keeps silent in pauses between sentences, such as the blue parts in Fig 3.1.

If the above-mentioned silence features can be learned, periods of silence can be detected, which means that periods of sentences can be detected, too. After that, the segmentation of speeches of sentences will become an easy job. In the following section, a method is proposed to detect the silence features, along with a sentence segmentation algorithm.

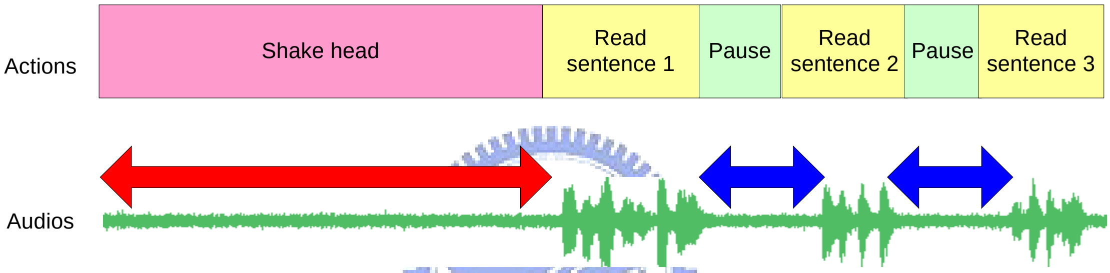


Fig. 3.1 A diagram that shows the audios, the corresponding actions taken, and the silence periods in a recorded video.

3.1.2.2 Segmentation Process

Before the segmentation can begin, the silence features must be learned first. To achieve this goal, the problem of the determination of "silence" must be solved. Silence here means audio parts recorded while the model does not speak. However, the volume of these parts usually is not zero due to the noise in the environment, so that they cannot be detected by simply searching zero-volume zones.

To decide a volume threshold for distinguishing silent parts from sound ones, the period when the model is shaking his/her head is utilized. Since the model is asked to keep silent in this period, the recorded volume originates from the noise in the environment. The maximum volume appeared in this period can be viewed as the threshold value. The duration of this period can be known easily because the system operator controls it.

After the threshold value is determined, the silent parts can be found by searching for the ones whose volumes are always smaller than the threshold value. However, short pauses between syllables in a sentence may be viewed as silences. To solve this problem, lengths of audio parts should be put into consideration, that is, the ones that are not long enough should be discarded. The duration of pauses between sentences are designed to be much longer than that of natural ones between syllables to avoid erroneous detections.

Finally, the silent audio parts can be found. Then, the sound parts can be found and segmented. The entire process of sentence segmentation is described as follows, and a flowchart of this process is shown in Fig. 3.2.

Algorithm 1. *Sentence segmentation by silence features.*

Input: an audio $A_{transcript}$ of the entire transcript, a predefined duration D_{shake} for shaking head, and a predefined duration D_{pause} for pausing between sentences.

Output: several audio parts of sentences $A_{sentence1}$, $A_{sentence2}$, etc.

Steps:

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

Step 2: Find a continuous audio part $A_{silence}$ whose volume is always smaller than V and lasts longer than D_{pause}.

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. 3.3 illustrates an example of the experimental results of the proposed segmentation method. The blue and green parts represent odd and even sentences, respectively. It is shown that the sound parts of the audio are learned correctly.

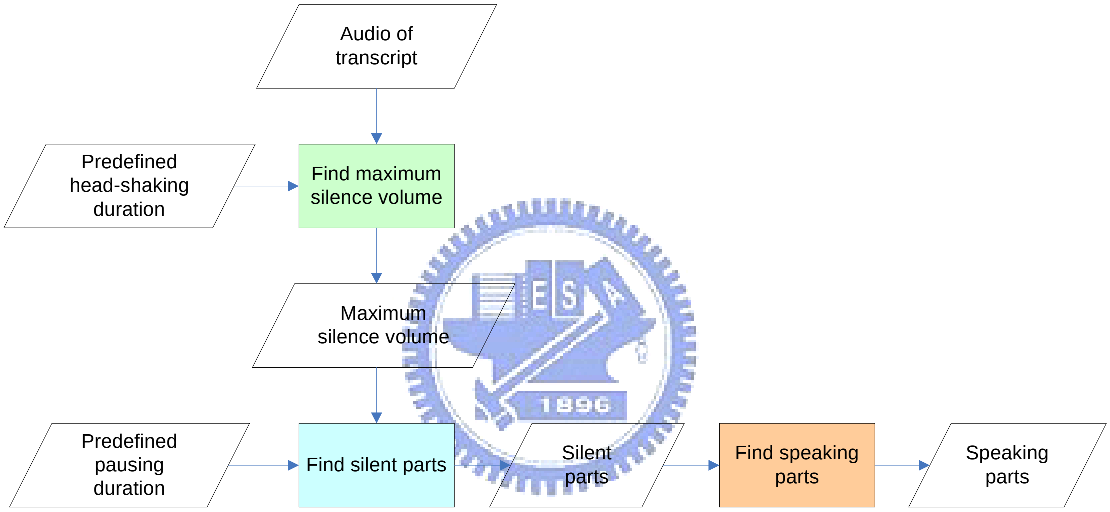


Fig. 3.2 Flowchart of the sentence segmentation process.

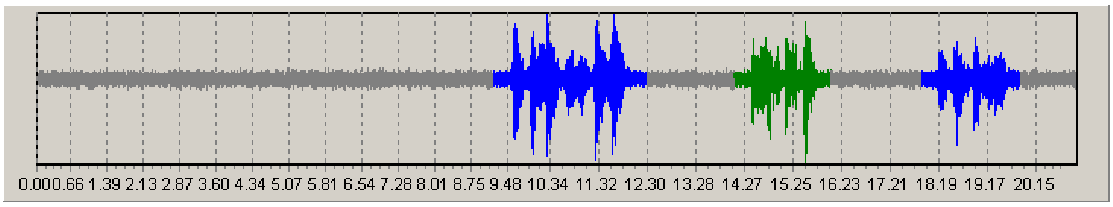


Fig. 3.3 An example of sentence segmentation results. The time of head shaking is 5 seconds, and the time of pausing between sentences is 1 second.

3.1.3 Review of Alignments of Mandarin Syllables

After the segmentation of sentences is done, the timing information of each syllable in a sentence can be learned by speech recognition or alignment techniques. The alignment ones are also kinds of speech recognition techniques, however, they need to know the syllables spoken in input speeches. Therefore, they produce recognition results with higher accuracy. In this study, a speech alignment technique using the Hidden Markov Model is utilized to learn the timing information of syllables.

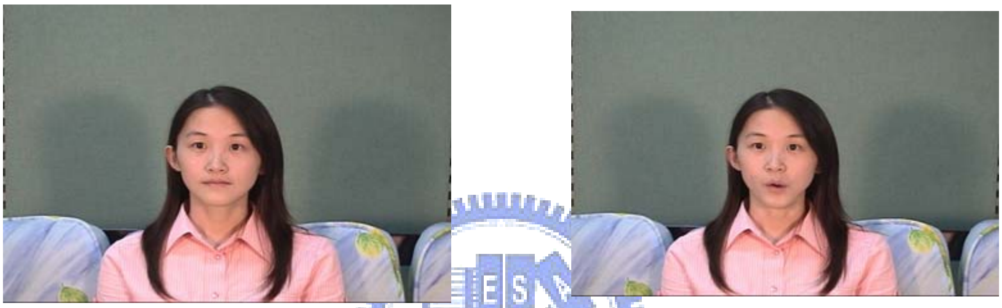
The Hidden Markov Model, which can be abbreviated as HMM, is a model for speech recognition and alignment using statistical methods. It is used to characterize the spectral properties of the frames of a speech pattern. In [6], Lin and Tsai adopted a sub-syllable model together with the HMM for recognition of Mandarin syllables. After the sub-syllable model is constructed, the Viterbi search is used to segment the utterance. Finally, the timing information of every syllable in the input speech is produced.

3.2 Learning of Base Image Sequences

As mentioned in Section 3.1, the silence period for the model to shake his/her head in the video recording process is used to help segment sentences. However, this period is designed to have another function, that is, to help learn base image sequences. In Section 3.2.1, the meaning and use of base image sequences is described. And in Section 3.2.2, a process that utilizes the silence period to learn base image sequences is proposed.

3.2.1 Descriptions

Base image sequences are sequences of base images, while a base image is a facial image onto which some mutable facial features may be pasted to form new images with different facial expressions. For instance, Fig. 3.4(a) shows an example of a base image. After pasting a new mouth image onto the correct position, a new face is produced, as illustrated by Fig. 3.4(b). In the same way, after pasting several new mouth images onto a sequence of base images, an animation of a talking face is produced.



(a) (b)

Fig. 3.4 Example of base images. (a) A base image. (b) A base image with a new mouth pasted on.

As mentioned above, base images provide places for variable facial features to be pasted on. These variable ones normally include eyebrows, eyes, mouths, etc. However, the mouths are the only kinds of features adopted as variable ones in this study. The eyebrows and eyes are not pasted onto base images; instead, the eyebrows and eyes on the base images are retained to produce animations with more natural eye-blinking actions.

The motion of a head is another kind of feature controlled by the base images. 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. In the same way, other kinds of head movements such as nodding can be integrated.

Base images control more things in the generated animation. For example, the background, the body, and the hand are all controlled by the base images.

3.2.2 Learning Process

To produce base image sequences, 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 shaking of heads while speaking, and the image frames recorded during this period are used as base images. Certainly, all image frames of the recorded video can be used as base images. However, there are some drawbacks.

Firstly, since the actions of eyes and eyebrows originate from the base images, namely, all the image frames of the recorded video, the model must keep his/her eyes "natural" during the entire recording process, which is a very tiring job. And secondly, since the model is asked to pause awhile between sentences, generated base image sequences will exhibit this behavior, which is somewhat unacceptable. To avoid these drawbacks, only the image frames recorded during the head-shaking period are used as base images in this study. The period is short and no pause exists in it.

To generate a sequence of base images, an initial frame is selected first. And then, a traverse direction, either forward or backward, is selected. Starting from the initial frame, the frames met along the traverse direction are added to the sequence. In order to create animations with changeful head motions, the initial frame and the traverse direction is randomly selected for every session of animation. Besides, since desired animations may require more image frames than the number of total base images, the images must 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. 3.5 illustrates this situation.

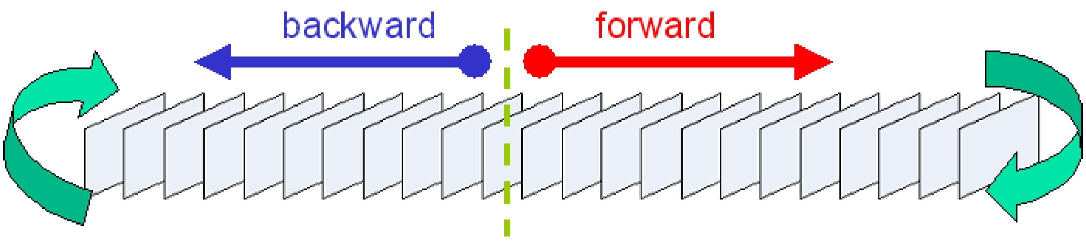


Fig. 3.5 A diagram that shows the generation process of base image sequences.

The entire process of generating base image sequences is described as follows, and a flowchart of this process is shown in Fig. 3.6.

Algorithm 2. *Learning process of base image sequences*.

Input: a sequence of image frames $I = \{I_1, I_2, \ldots, I_M\}$ of the recorded video in the head-shaking period, and the amount of desired base images N.

Output: a sequence of base images $B = \{B_1, B_2, \ldots, B_N\}$.

Steps:

Step 1: Randomly select an initial frame $I_{initial}$ in I.

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

Step 3: Add the current frame to B.

Step 4: Stop learning, if the number of frames in B equals to N.

Step 5: Reverse the direction if the current frame is I_1 or I_M.

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

Step 7: Repeat Steps 3 through 6.

3.3 Experimental Results

In this section, some experimental results of the proposed methods described in this chapter are shown. Firstly, Fig. 3.7 shows the entire audio of a transcript, and Fig. 3.8 shows the sentence segmentation result. The 17 speaking parts of the audio, which are represented in blue and green colors, are detected successfully.

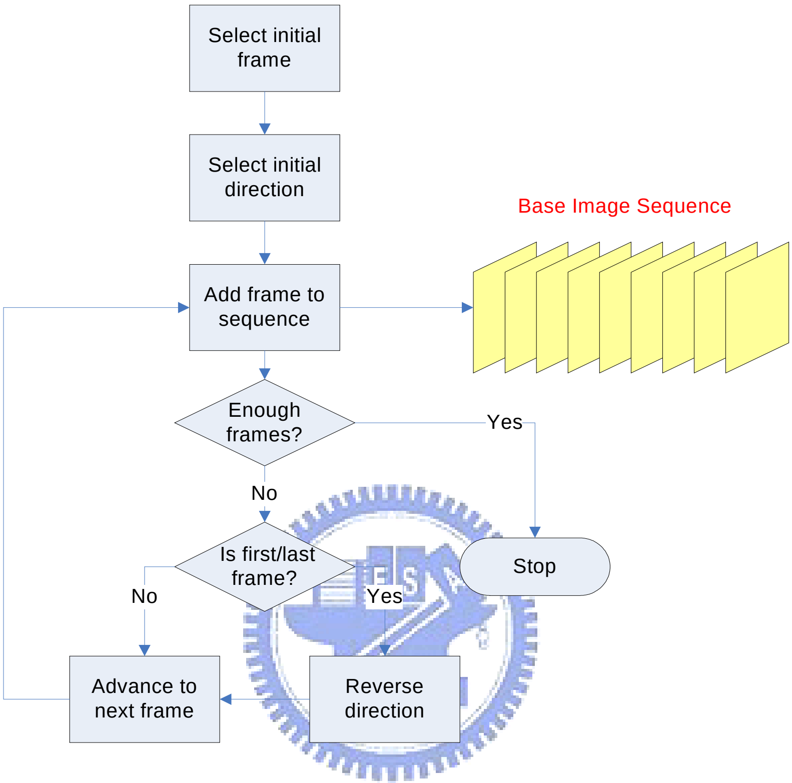


Fig. 3.6 Flowchart of the learning process of base image sequences.

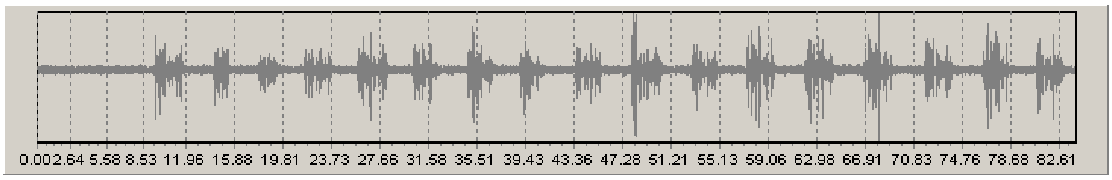


Fig. 3.7 An example of entire audio data of a transcript. The duration of head shaking is 5 seconds, and the duration of pausing between sentences is 1 second.

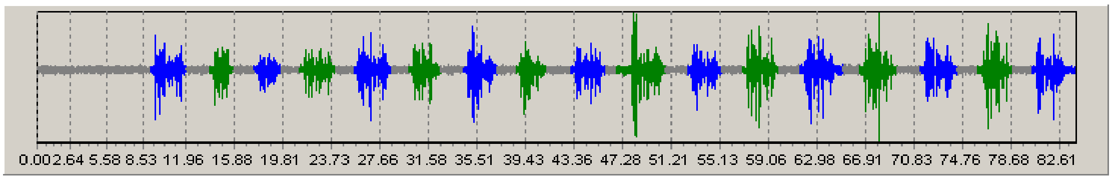


Fig. 3.8 The audio data in Fig. 3.7 is segmented into 17 parts.

Secondly, Fig. 3.9 shows an audio containing a Mandarin sentence, and Fig. 3.10 shows the result of syllable alignment. Durations of syllables are shown in blue and green colors. In Fig. 3.11 and 3.12, another Mandarin sentence and its corresponding result of syllable alignment are shown.

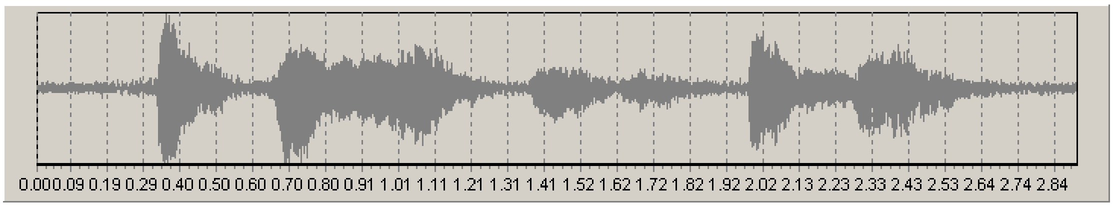


Fig. 3.9 An audio that contains a Mandarin sentence “好朋友自遠方來”.

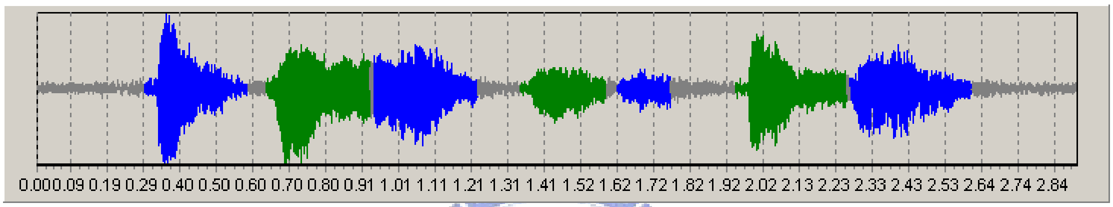


Fig. 3.10 The result of syllable alignment of the audio in Fig. 3.9.

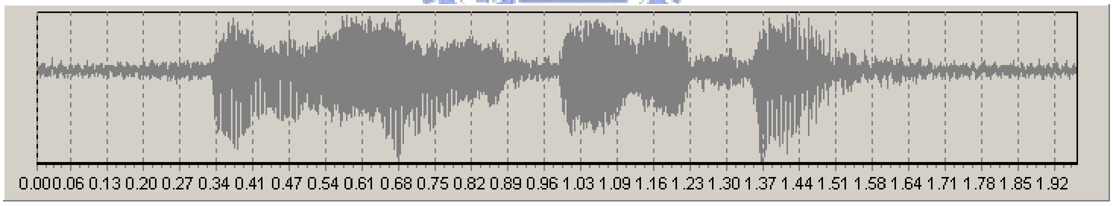


Fig. 3.11 An audio that contains a Mandarin sentence “熟能生巧”.

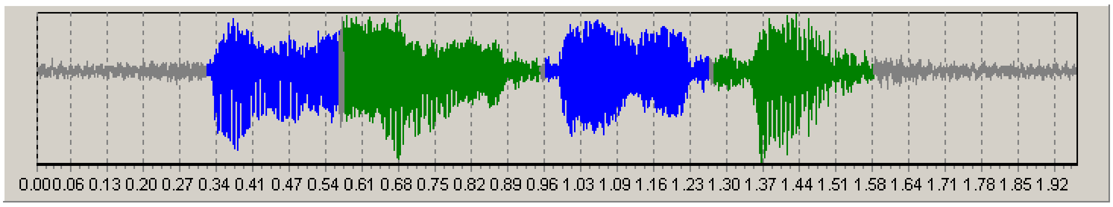


Fig. 3.12 The result of syllable alignment of the audio in Fig. 3.11.

Thirdly, Fig. 3.13 shows a base image sequence produced with the proposed method.

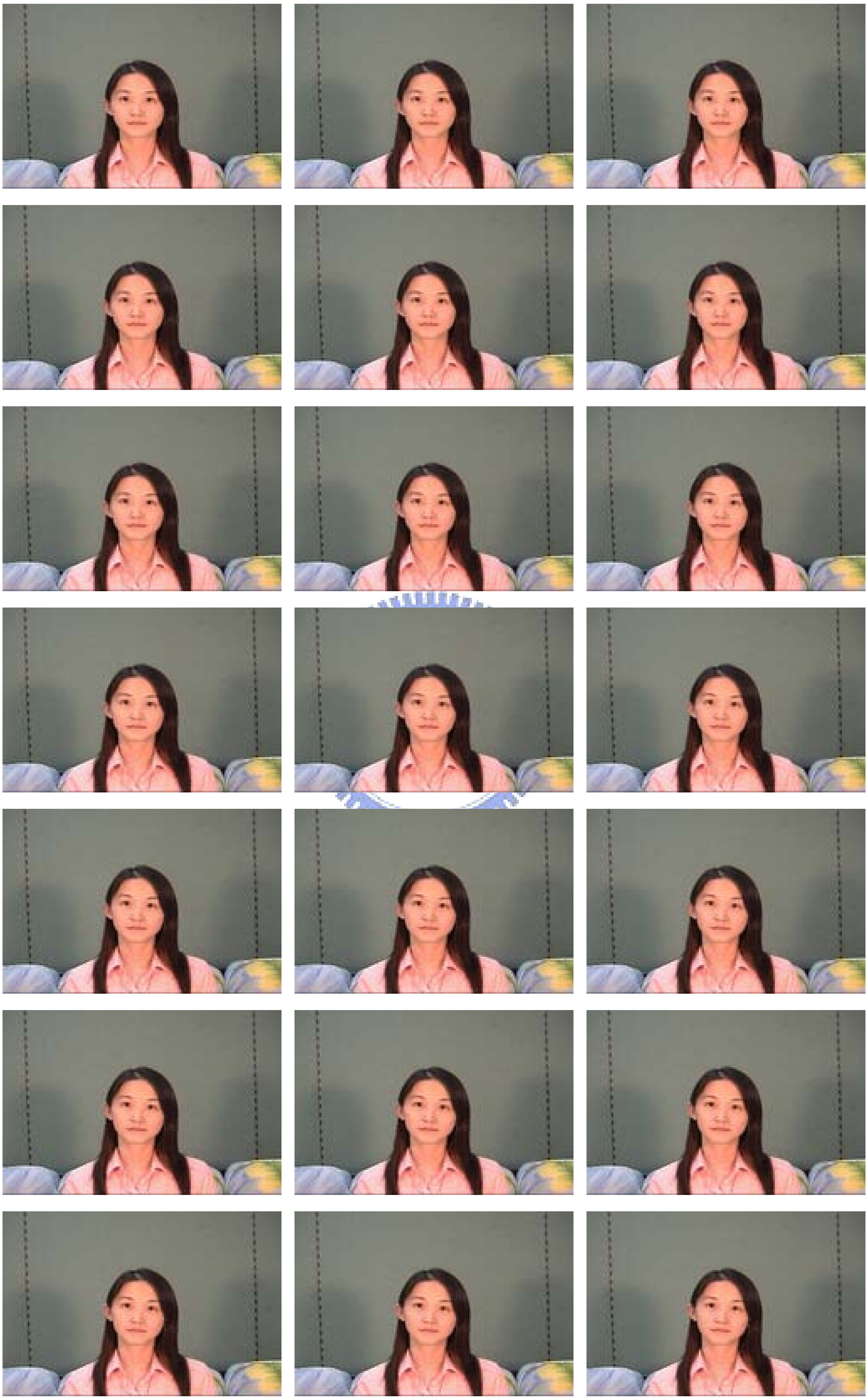


Fig. 3.13 A base image sequence produced with the proposed method.

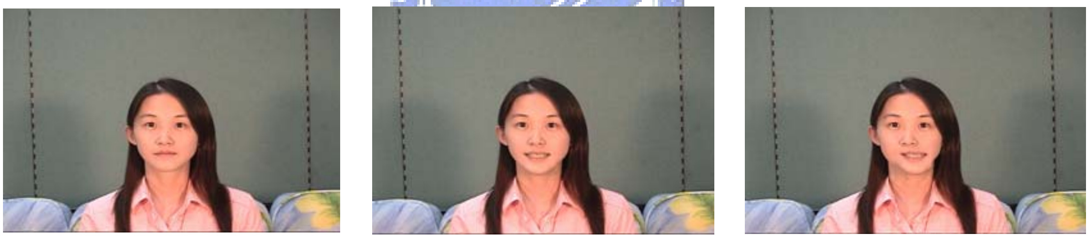
3.4 Summary and Discussions

In this chapter, an automatic method for sentence segmentation is proposed. The method works well in silent environments. The method is also workable in environments with constant noise, such as the noise of fans and cooling systems. The segmentation helps accelerate the subsequent work of syllable alignment. Besides, a method for generating base image sequences by utilizing the period of head shaking is proposed. The sequences generated are different for every session, which helps prepare varied background images for animations.

Chapter 4
Automatic Learning of Facial Features

4.1 Introduction

To create an animation of a speaking person, syllables spoken are collected first, and then visemes corresponding to the syllables must be “pasted” onto the base image sequence. The visemes, namely, the mouth images, should be pasted onto correct positions of faces; otherwise the generated animation will look strange. As shown in Fig. 4.1, pasting on incorrect positions leads to unacceptable results.



(a) (b) (c)

Fig. 4.1 Example of base images. (a) A base image. (b) The base image with a new mouth pasted onto the correct position. (c) The base image with a new mouth pasted onto an incorrect position.

In order to decide the correct positions for the mouth images to be pasted on, three types of methods have been tried in this study. The first one is to measure the positions manually. Obviously, the positions obtained may be very precise. However, it is not suitable to perform this work on many frames. The second is to plaster the face with some marks, and then the positions can be detected easily and automatically.

However, this method bears the disadvantage of plastering extra marks on the face. The third method is to measure the positions by face recognition techniques on every frame. This method is fully automatic, however, results of recognition are often not stable enough due to slight variations in lighting. The slight movements of muscles under the skin also may affect the recognition results significantly, though human eyes may not notice them.

In this study, a method that integrates the second and the third method mentioned above is proposed. A face recognition technique using a knowledge-based approach is used to learn the positions of facial features for the first frame. The technique is reviewed in Section 4.2. Spatial relations between these features, which keep invariable for an identical face, are noted. Then, a kind of facial feature is used as a sort of mark, and this “natural” mark, which is called the base region in this study, can be detected by image-matching techniques. Finally, the positions of other facial features excluding the mark can be calculated according to the spatial relations. The process is illustrated in Fig. 4.2.

One advantage of this method is that the results of matching are more stable while the mark keeps unchanged for every frame. Another advantage is that the image matching techniques can even be applied to rotated faces, which is discussed in Section 4.3.2.

To select a proper facial feature to be used as the base region, its invariance is important. Among those facial features listed in Fig. 4.3, the nose is the only one that keeps an invariant shape while the face is speaking. The eyebrows may move slightly due to expressions and the eyes may blink casually. The shapes of the mouth and the jaw change obviously on a speaking face. Therefore, the nose is selected as the base region in this study.

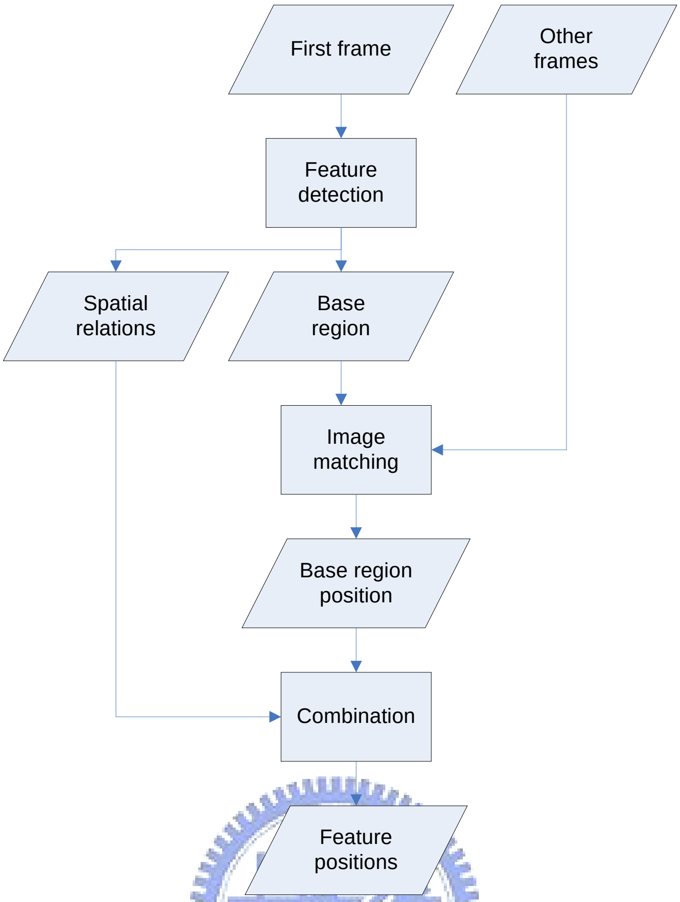
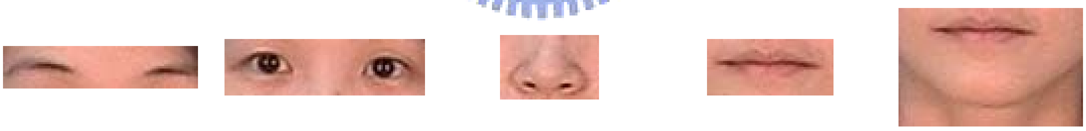


Fig. 4.2 Flowchart of the learning process of facial features.



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

Fig. 4.3 Facial features. (a) The eyebrows. (b) The eyes. (c) The nose. (d) The mouth. (e) The jaw.

4.2 Review of Knowledge-Based Face Recognition Techniques

Knowledge-based face recognition techniques use the common knowledge of facial features to detect their positions. An example of the knowledge is that eyes on a face have similar shapes. Another example is that eyebrows have similar shapes while they are always above the eyes.

In this study, relations and shapes of facial feature are used as the knowledge to learn their positions. First, the skin part of a facial image is found by color thresholding. Facial features are filtered according to the feature properties and relations. Edges of the image are used to find the positions of the facial features more precisely.

4.3 Learning of Base Regions

In this section, the proposed learning process of base regions is described in detail. After the position of the base region of the first frame is determined using the technique described in Section 4.2, the process is performed on other frames to learn the positions of the base regions of them. The positions of facial features can then be determined easily.

4.3.1 Review of Image Matching Techniques

In Section 4.1, it is mentioned that the base region positions of the frames other than the first one can be determined using image matching techniques. These techniques are used to find the position of a pattern image inside a base image. Fig. 4.4 shows the block diagram of common image matching techniques. By shifting the position of the pattern image (or the base image on the contrary), several measures can be calculated according to a pre-designed formula. At last, the position corresponding to the best measure is adopted.

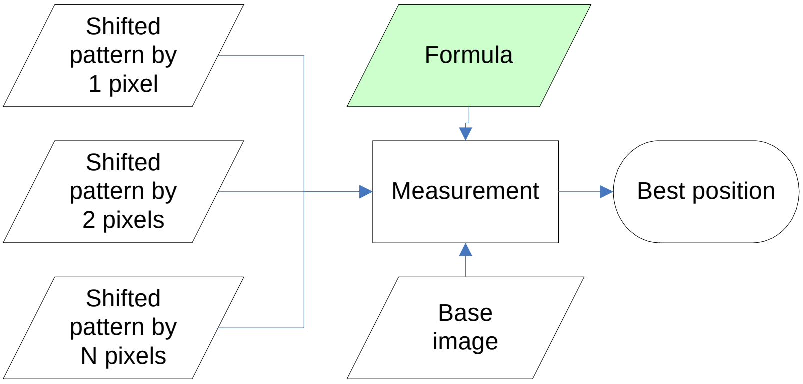
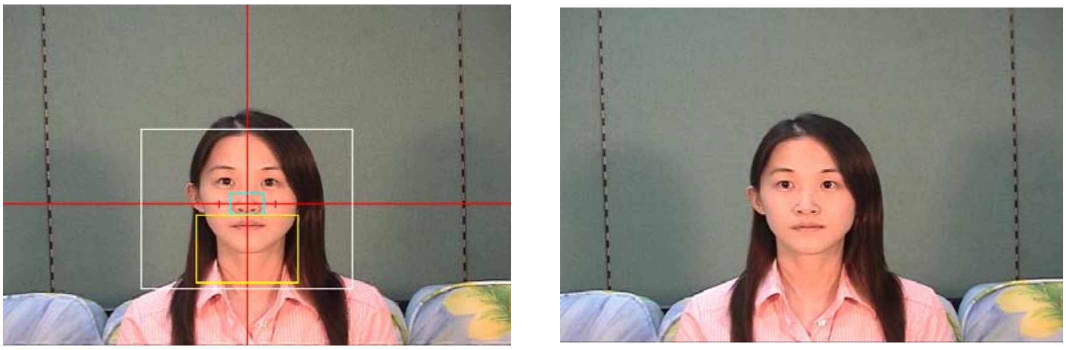

Shifted pattern by 1 pixel
Shifted pattern by 2 pixels
Shifted pattern by N pixels
Formula
Measurement
Best position
Base image

Fig. 4.4 Block diagram of common image matching techniques.

It is obvious that the formula affect the results severely. In an environment with controlled lighting, a formula that calculates the “Euclidean distance” of two images is sufficient. For example, Equation (4.1) below is used in this study to calculate the Euclidean distance between the colors of two images:

$$D = \sum_{x}\sum_{y}(R_1(x,y) - R_2(x,y)) + (G_1(x,y) - G_2(x,y)) + (B_1(x,y) - B_2(x,y)) \quad (4.1)$$

An example of results of image matching using Equation (4.1) is illustrated in Fig. 4.5. Fig. 4.5(a) shows the first frame of a recorded video, and the blue block on it is the base region detected using the techniques described in Section 4.2. Fig. 4.5(b) shows the 2617th frame of the video, and the base region position of it is calculated using the image matching technique mentioned above.



(a) (b)

Fig. 4.5 Example of image matching results. (a) The first frame of a video. The base region position is (346, 280). (b) The 2617th frame of the video. The base region position is (353, 283).

4.3.2 Learning by Image Matching with Sub-Pixel Precision

The image matching technique used in Section 4.3.1 is proper for finding the base region positions of a face, even if the face is shaking. Table 4.1 shows an example of base region positions of a sequence of images. It is shown that the minimum unit of a base region position is a pixel. However, the face normally does not shake in the way of shifting its position by one pixel for successive frames suddenly, instead, smoothly. Therefore, to find the positions with sub-pixel precision is useful for high-quality animations. Examples of positions with sub-pixel precision are (346.5, 280.5) and (353.3, 283.6).

To find a position with sub-pixel precision, a pattern image needs to be shifted by a distance shorter than a pixel, and then the image matching technique can be performed on the shifted pattern image. To accomplish the job of shifting, the continuous properties of facial images are utilized.

In [8], Gonzalez and Woods illustrated the process of acquiring digital images from sensors. The image acquired from the sensors is continuous with respect to the x- and y-coordinates, and also in amplitude. The coordinate values and amplitude values are sampled and quantized into digital forms, respectively. The situation is shown in Fig. 4.6. Pink arrows in Fig. 4.6 indicate positions between pixels. If the amplitude values of these positions can be known, the image matching technique can be performed on these values, just like the image is shifted within a pixel.

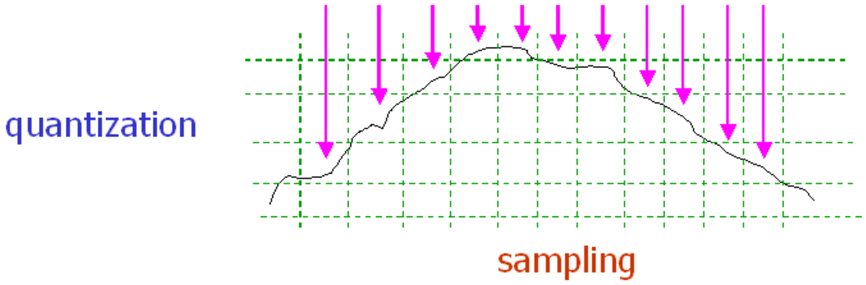


Fig. 4.6 A diagram of converting a continuous image into a digitized form.

Since the face images are continuous, it is reasonable to assume that the amplitude values, namely, the color values, between two adjacent pixels approximate the values of these two pixels. In this study, the technique of bilinear interpolation is used to generate these values. Fig. 4.7 illustrates this technique. The color value of P’ is determined in proposition to the color values of the nearest four pixels P_1 through P_4 using following equations:

$$A_1 = |(x' - x_1)(y' - y_1)|,\ A_2 = |(x' - x_2)(y' - y_2)|;$$

$$A_3 = |(x' - x_3)(y' - y_3)|,\ A_4 = |(x' - x_4)(y' - y_4)|; \tag{4.2}$$

$$P' = (A_1P_4 + A_2P_3 + A_3P_2 + A_4P_1) / (A_1 + A_2 + A_3 + A_4). \tag{4.3}$$

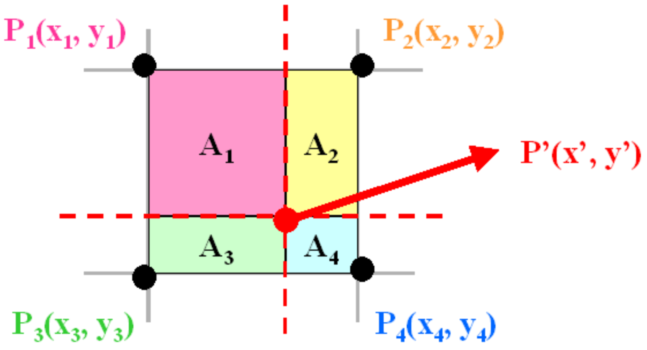


Fig. 4.7 A diagram of the adopted bilinear interpolation technique.

A method that performs image matching with sub-pixel precision using above-mentioned ideas and techniques is proposed and described as follows. Firstly, the pattern image, namely, the image of the base region, and the base image are enlarged with a predefined ratio using the bilinear interpolation technique. The color values of pixels that have no corresponding pixels in the original image are filled with interpolated values. Secondly, the image matching technique described in Section 4.3.1 is applied on the enlarged pattern and the base image to find the best position of the base region. Finally, the position is shrunk back according to the predefined ratio.

The algorithm of the image matching with sub-pixel precision is described as follows.

Algorithm 3. *Image matching with Sub-Pixel precision.*

Input: a pattern image $I_{pattern}$, a base image I_{base}, and a predefined ratio r.

Output: the position $P(x, y)$ of $I_{pattern}$ in I_{base}.

Steps:

Step 1: Enlarge $I_{pattern}$ r times larger to get a new image $I_{patternL}$.

Step 2: Enlarge I_{base} r times larger 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}.

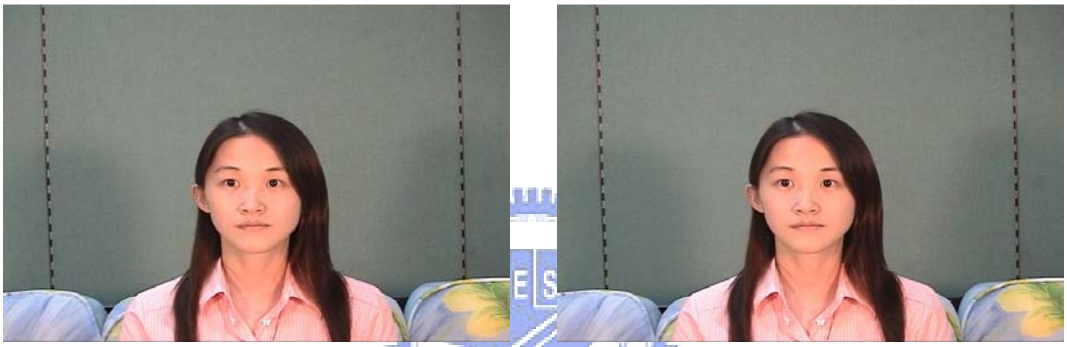
Step 4: Divide x' by r to get x.

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

An example of results of the proposed method is shown in Fig. 4.8. The ratio value is 2 in this example.

4.3.3 Handling Rotated Faces

In the video recording process, a model is asked to shake his/her head for a period of time, so that generated base image sequences can exhibit a speaking face with natural shaking. Besides, the face of the model may not always be straight due to his/her speaking habit. However, the image matching technique cannot be applied effectively on those shaking faces because the base regions are “rotated”. In Section 4.3.3.1, problems caused by rotated faces are described. In Section 4.3.3.2, a modified image matching method that is suitable to be applied to the rotated faces is proposed.



(a) (b)

Fig. 4.8 Example of results of image matching with sub-pixel precision. (a) The 2597th frame of a video. The base region position is (354.5, 284.0). (b) The 2598th frame of the video. The base region position is (354.5, 284.5).

4.3.3.1 Problems of Rotated Faces

The image matching technique described in Section 4.3.1 is effective to find the position of a pattern image in a base image. However, this is true only under the assumption that the pattern part of the base image is very similar to the pattern image. For rotated faces, the pattern parts, namely, the base regions in this study, are rather different from the base region of the first frame. Therefore, results of image matching are often not quite precise.

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. Fig. 4.9 shows the problem. As described in Section 4.1, positions of facial features are calculated according to the base region position and spatial relations. For a straight face like Fig. 4.9(a), the positions of 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. To determine the positions of facial features correctly on a rotated face, the rotated angle must be found.

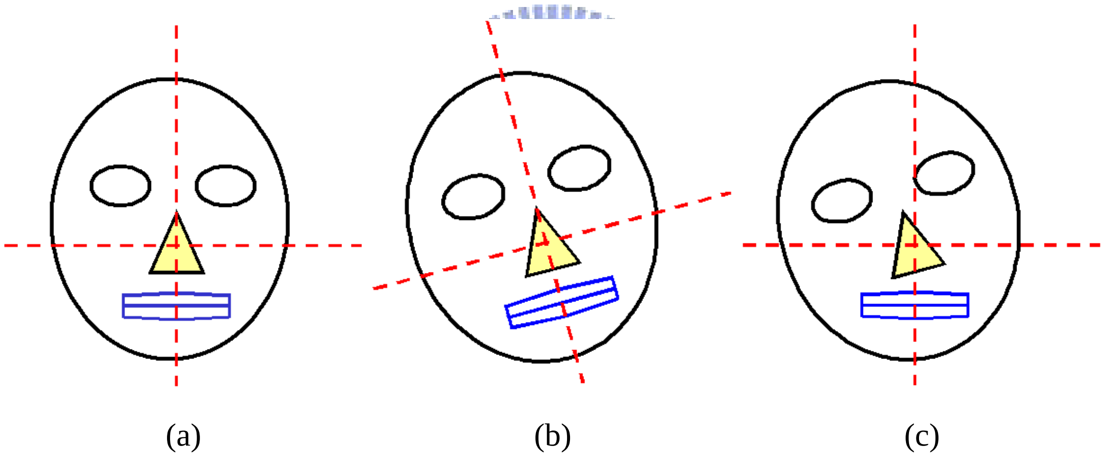


Fig. 4.9 A diagram that shows the problem of rotated faces. (a) A straight face with its mouth position determined correctly by the spatial relation. (b) A rotated face. (c) A rotated face with its mouth position determined incorrectly by the spatial relation.

4.3.3.2 Image Matching on Rotated Faces

To find the rotated angle of a rotated face, some extra work is added to the original image matching method. As shown in Fig. 4.10, the base region image is rotated first to generate several rotated versions. All of these rotated images as used as pattern images. And then, the image matching technique is applied. Finally, the best position and rotated angle of the base region is learned. The algorithm of this process is described as follows.

Algorithm 3. *Image matching on rotated faces*.

Input: a pattern image $I_{pattern}$, and a base image I_{base}.

Output: the position and rotated angle of $I_{pattern}$ in I_{base}.

Steps:

Step 1: Rotate $I_{pattern}$ with an incremental series of degrees and get a set of new images $I_{patternR}$.= $\{I_{patternR1}, I_{patternR2}, \ldots, I_{patternRN}\}$.

Step 2: Select an image $I_{patternR'}$ in $I_{patternR}$ as the pattern image.

Step 3: Perform image matching on $I_{patternR'}$ and I_{base} and record the measurement value.

Step 4: Repeat steps 2 through 3 until all images in $I_{patternR}$ are used ever.

Step 5: Output the position and rotated angle corresponding to the minimum measure value.

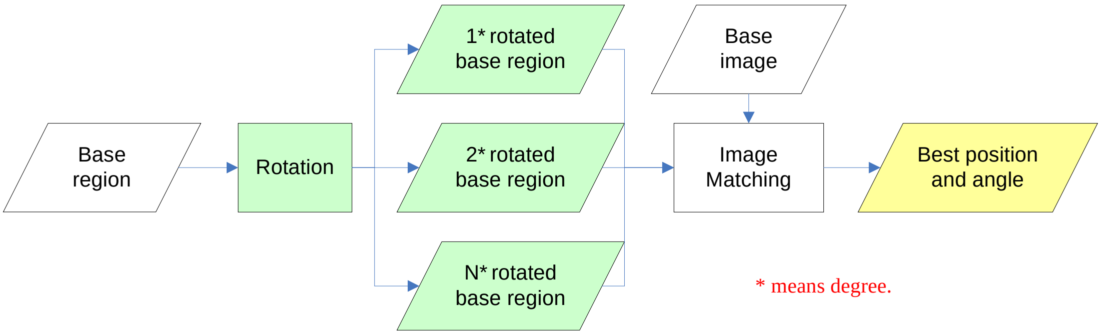
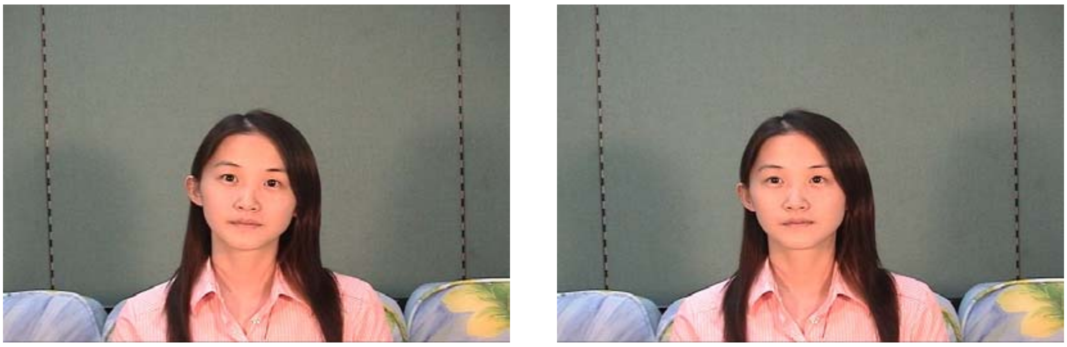


Fig. 4.10 Flowchart of the proposed image matching method on rotated faces.

An example of results of image matching on rotated faces is shown in Fig. 4.11.



(a) (b)

Fig. 4.11 Example of results of image matching on rotated faces. (a) The rotated angle is 6 degrees clockwise. (b) The rotated angle is –1 degree clockwise.

4.3.4 Learning by Image Matching with Sub-Pixel Precision on Rotated Faces

The techniques mentioned in Sections 4.3.2 and 4.3.3 can be combined together to learn the positions and rotated angles on rotated faces with sub-pixel precision. Fig. 4.12 illustrates the combined process, in which green blocks represent the work of matching on rotated faces, and blue blocks represent the work of matching with sub-pixel precision.

4.3.5 Correcting Erroneous Matching Results

Although the results of the image matching technique are stable, however, some errors still exist due to unavoidable changes in lighting. The variations of shadows on faces also affect the accuracy of the results.

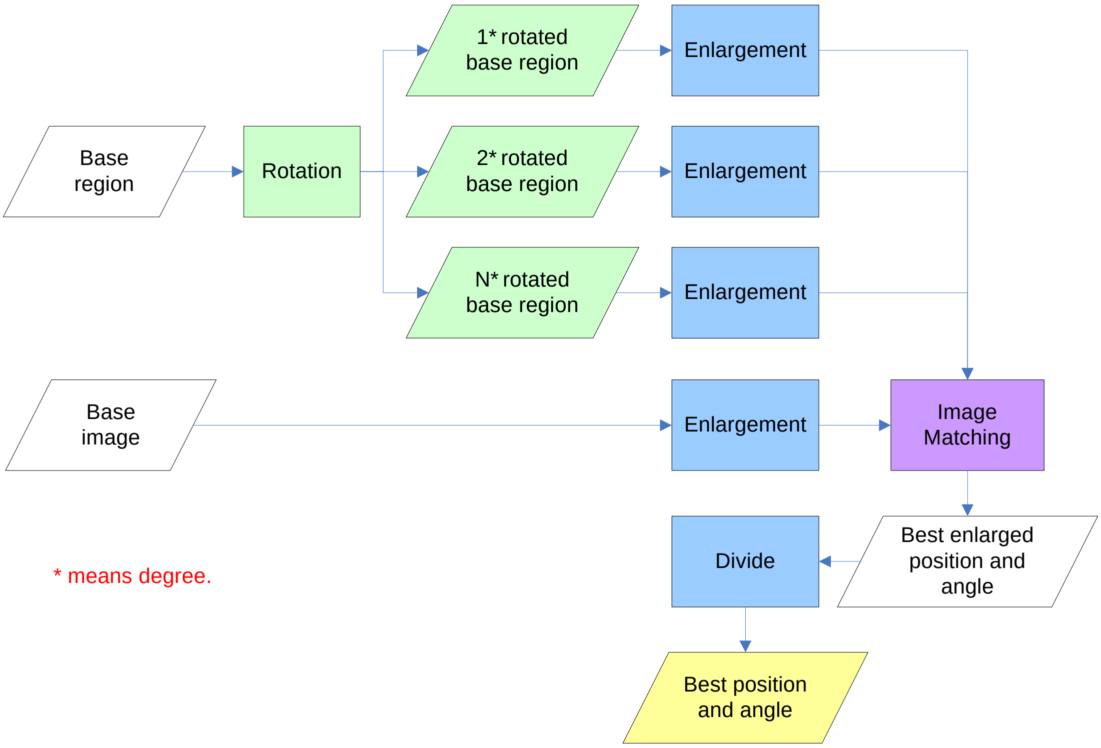


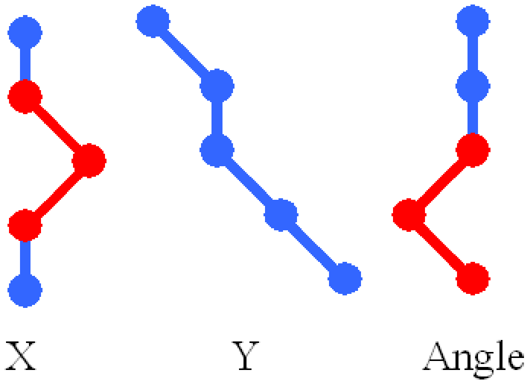
Fig. 4.12 Flowchart of image matching with sub-pixel precision on rotated faces.

For instance, Fig. 4.13(a) shows an example of the result obtained using the proposed method in Section 4.3.4. Corresponding trajectories of the x-axis, the y-axis and the rotated angle are shown in Fig. 4.13(b). Red parts of the trajectories represent situations that the position of the base region suddenly moves forward along a direction and then back. For example, the x-axis of the base region of frame 2 is 344.0, and then it moves right to 344.5 on frame 3, and then it moves left to 344.0 on frame 4. This kind of situation is not normal since the face often shakes smoothly.

To solve this problem, a method is proposed to correct the erroneous values. It simply deals with the following situation: the position of the base region suddenly moves forward along a direction and then back in three frames. In this situation, the base region position of the 2nd frame is corrected with the center of the base regions positions of the 1st and the 3rd frame.

Frame	X	Y	Angle
1	344.0	280.0	0
2	344.0	280.5	0
3	344.5	280.5	0
4	344.0	281.0	1
5	344.0	281.5	0

(a)



(b)

Fig. 4.13 Example of erroneous matching results. (a) The x- and y-axis and rotated angle detected by image matching techniques. (b) The trajectories of the x-axis, the y-axis, and the rotated angle.

The entire algorithm of correcting erroneous matching results is described as follows:

Algorithm 4. *Correcting erroneous matching results.*

Input: a sequence of triples (x_i, y_i, A_i) = { (x_1, y_1, A_1), (x_2, y_2, A_2), …, (x_N, y_N, A_N)}, where x_i, y_i, and A_i represent the x-axis, the y-axis and the rotated angle of the base region of frame *i*, respectively.

Output: a sequence of corrected triples (x_i, y_i, A_i) = { (x_1, y_1, A_1), (x_2, y_2, A_2), …, (x_N, y_N, A_N)}.

Steps:

Step 1: Set *current* to 1.

Step 2: Set $x_{current}$ to $(x_{current\text{-}1}+x_{current+1})/2$, if $(x_{current}-x_{current\text{-}1})(x_{current}-x_{current+1}) > 0$.

Step 3: Set $y_{current}$ to $(y_{current\text{-}1}+y_{current+1})/2$, if $(y_{current}-y_{current\text{-}1})(y_{current}-$

$y_{current+1}$) > 0.

Step 4: Set $A_{current}$ to ($A_{current\text{-}1}+A_{current+1}$)/2, if ($A_{current}-A_{current\text{-}1}$)($A_{current}-A_{current+1}$) > 0.

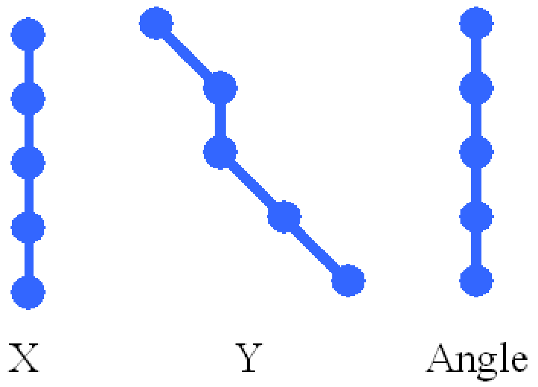
Step 5: Add *current* by 1.

Step 6: Repeat Steps 1 through 5 until *current* is larger than *N*.

Fig. 4.14 shows the result of Fig. 4.13(a) after correcting erroneous matching results. The trajectories shown in Fig. 4.14(b) are smoother and more natural.

Frame	X	Y	Angle
1	344.0	280.0	0
2	344.0	280.5	0
3	344.0	280.5	0
4	344.0	281.0	0
5	344.0	281.5	0

(a)



(b)

Fig. 4.14 Example of corrected matching results. (a) The values in Fig. 4.13(a) are corrected. (b) The corrected trajectories of the x-axis, the y-axis, and the rotated angle.

4.4 Experimental Results

In this section, some experimental results of the proposed methods described in this chapter are shown. Fig. 4.15 shows a sequence of image frames in a recorded video. Positions and rotated angles of base regions detected using the method proposed in Section 4.3.4 are listed in Table 4.1(a), and the corrected values using the method proposed in Section 4.3.5 are listed in Table 4.1(b).

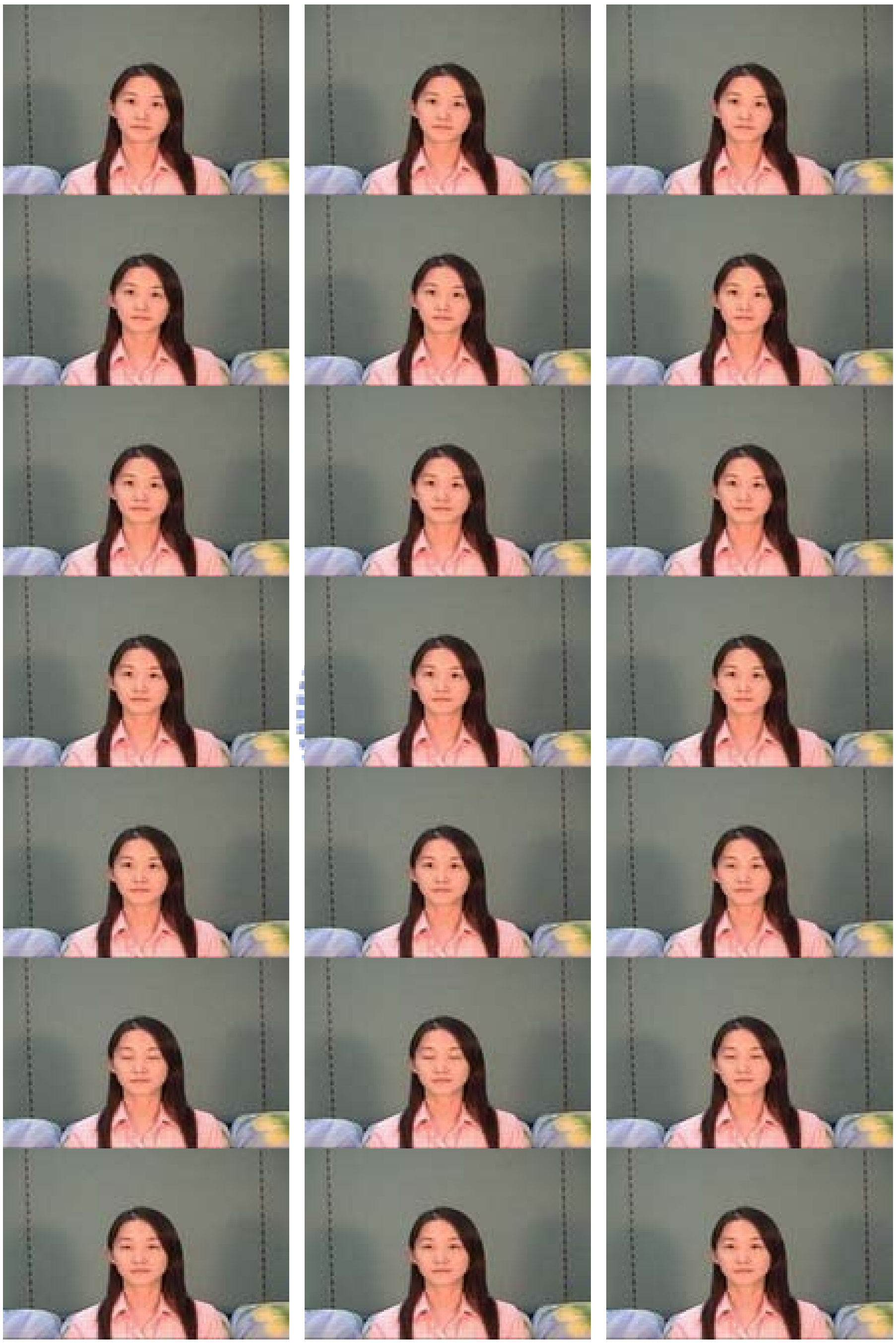


Fig. 4.15 A sequence of image frames in a recorded video.

Table 4.1 Positions and rotated angles of base regions of frames in Fig. 4.15. (a) Uncorrected values. (b) Corrected values.

Frame	X	Y	Angle
1	347.5	279.5	5
2	347.5	279.5	5
3	347.0	279.5	5
4	346.5	279.0	4
5	346.0	279.0	4
6	345.5	279.0	4
7	345.0	279.0	4
8	344.5	279.0	4
9	343.5	279.5	3
10	343.0	279.5	3
11	343.0	279.5	2
12	342.5	279.5	2
13	342.5	279.5	2
14	342.0	279.5	1
15	342.5	280.0	0
16	342.5	280.0	0
17	342.0	279.5	0
18	341.0	278.5	-1
19	341.5	279.0	0
20	340.5	278.0	-1
21	341.0	278.5	0

(a)

Frame	X	Y	Angle
1	347.5	279.5	5
2	347.5	279.5	5
3	347.0	279.5	5
4	346.5	279.0	4
5	346.0	279.0	4
6	345.5	279.0	4
7	345.0	279.0	4
8	344.5	279.0	4
9	343.5	279.5	3
10	343.0	279.5	3
11	343.0	279.5	2
12	342.5	279.5	2
13	342.5	279.5	2
14	342.5	279.5	1
15	342.5	280.0	0
16	342.5	280.0	0
17	342.0	279.5	0
18	341.75	279.25	0
19	341.5	279.0	0
20	340.25	278.75	0
21	341.0	278.5	0

(b)

4.5 Summary and Discussions

In this chapter, the automation of facial feature learning was emphasized. Since the positions of these features can be calculated according to the position of the base region, detection of the base region automatically becomes very important. The image matching technique may be used to do this job. Proposed methods modify the technique so that it can be performed on rotated faces with sub-pixel accuracy. A method is also proposed to correct erroneous matching results.

Chapter 5
Virtual Talking Face Animation Generation

5.1 Introduction

In Chapters 3 and 4, the audio features, the base image sequences, and the facial features are collected using the proposed methods. With the help of these features, virtual talking face animations with synchronized utterances can be generated. Fig. 5.1 shows a block diagram for the proposed animation generation process. “Proper” frames are generated according to the timing information of the input audio and the feature information in the viseme database. Finally the input audio and the generated frames are combined together to produce an animation. It is obvious that generating “proper” frames is a very important work. Badly generated frames may lead to asynchronous problems between the audio and the frames, or unnatural animations.

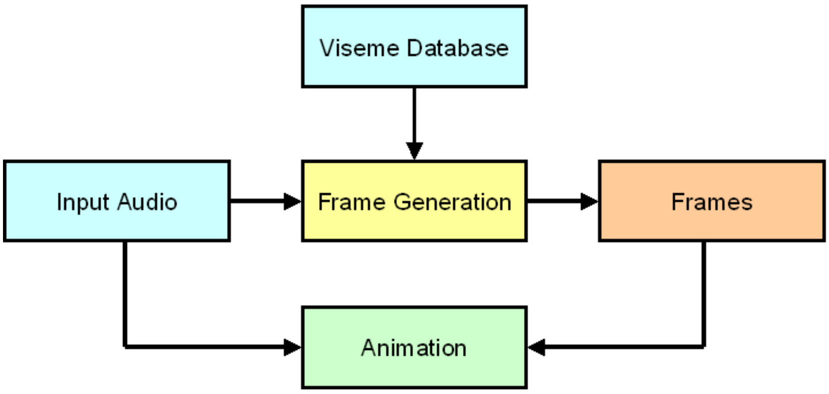


Fig. 5.1 Block diagram of proposed animation generation process.

Fig. 5.2 illustrates the frame generation process proposed by Lin and Tsai [6].

Firstly the timing information of the syllables of an input speech is obtained. This timing information is used to decide the number of frames to preserve every syllable and the pauses between syllables. The starting frame for every syllable is calculated by accumulating the number of frames of preceding syllables and pauses. Secondly, the corresponding visemes of syllables are prepared. Since the number of frames of a syllable in the viseme database may not equal the number of frames preserved, an algorithm for frame increase and decrease is used. Thirdly, the mouth images of visemes are pasted onto the base images. Finally, transition frames between syllables are replaced with several middle frames to produce smoother animations.

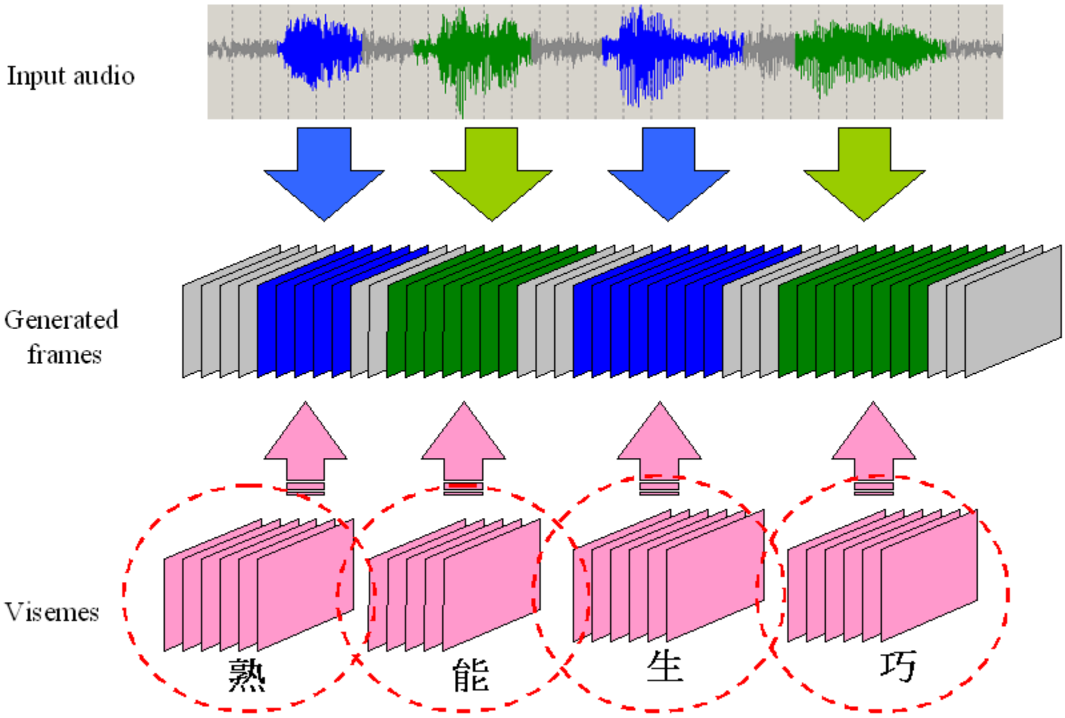


Fig. 5.2 Diagram of frame generation process proposed by Lin and Tsai [6].

The above-mentioned process is effective in generating proper frames for synchronized and smooth animations. However, some problems still exist. In the following sections of this chapter, the problems are illustrated, and some methods are proposed to solve these problems in order to produce higher-quality animations.

5.2 Synchronization Between Frames And Speeches

In Lin and Tsai [6], the starting frame for every syllable is calculated by accumulating the number of frames of the preceding syllables and pauses. To obtain the number of frames of a syllable or pause, the length of it is multiplied by a frames-per-second constant. The constant controls the number of frames appearing within a second. For a standard NTSC video, the constant is set to 29.97.

Since the length of a syllable or a pause is a floating-point number, the calculated number of frames is also a floating-point number. However, the number of frames can only be an integer. Discarding the fraction part leads to a small error. Errors are accumulated and affect the synchronization between the audio and the frames.

To solve this problem, a method of re-synchronization on every syllable is proposed. Firstly, the starting frame of every syllable is calculated in floating-point number, like F_0 through F_4 in Fig. 5.3. Then the fraction part of these floating-point frames are discarded, therefore results in starting frames of the integer type, like I_0 through I_4. The starting frames in the integer type are used as the starting frames of syllables. The number of frames of a syllable is obtained by subtracting the nearest successive starting frames of the integer type. For example, the number of the first syllable (the first green part) is $I_1 - I_0$, and the number of frames for the first pause between syllables (the first blue part) is $I_2 - I_1$.

The final result of duration of the integer type can be used to generate animations synchronized with the input audio, with at most an error of a frame.

5.3 Frame Generation by Interpolation

In the following two sections, two processes of frame generation for two applications are proposed. In Section 5.3.1, a frame generation process that is suitable for creating virtual talking faces is proposed. And in Section 5.3.2, a frame generation process that is suitable for creating virtual singing faces is described.

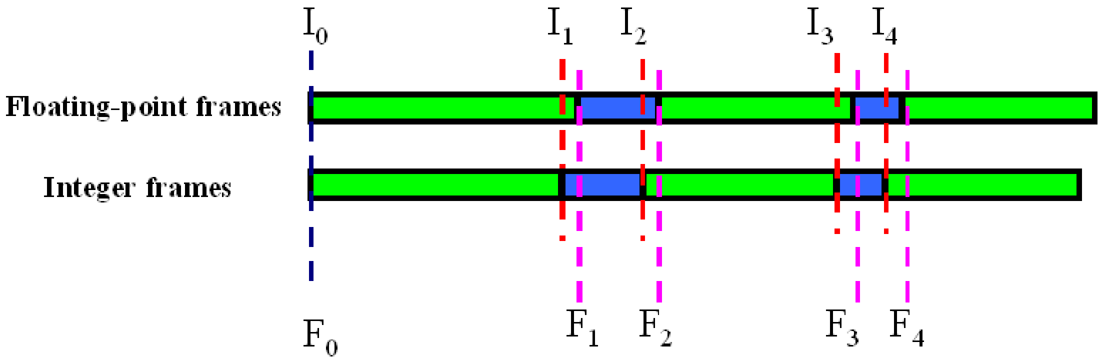


Fig. 5.3 Diagram of re-synchronization on every syllable.

5.3.1 Speaking Case

In Section 5.2, the number of frames for every syllable in an input audio is determined. The number of frames cannot be altered, or the asynchronous problem between the audio and image frames will arise. However, the number of frames for an identical syllable in the viseme database may not equal the one desired due to different speaking speeds and must be altered.

In [6], Lin and Tsai proposed an algorithm for frame increase and decrease to complete this job. The algorithm inserts frames at positions where adjacent frames are mostly unlike, and deletes frames at positions where adjacent frames are mostly alike, until the number of frames is equal to the desired one. However, results produced by this algorithm are not reasonable. It is noticed that when a person speaks faster, the shape of his/her mouth changes more violent. On the contrary, when a person speaks slower, the shape of his/her mouth changes slighter. However, the motion of the mouth retains the same when speaking an identical syllable.

To simulate the motion mentioned above, the idea of interpolation is used. As shown in Fig. 5.4, 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.

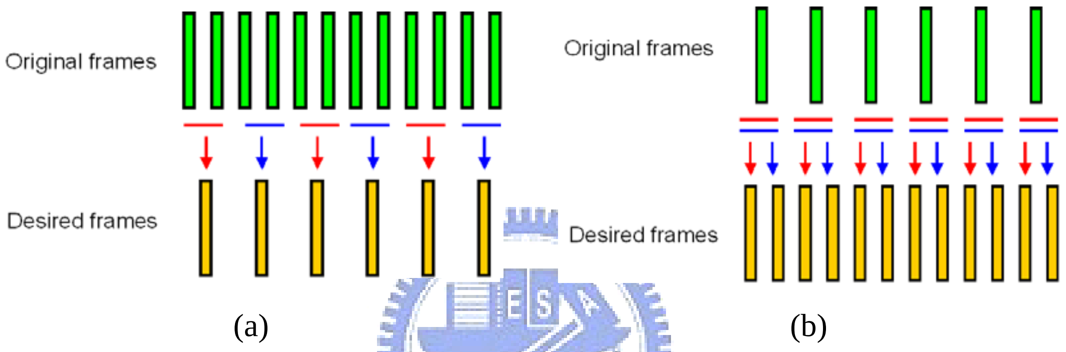


Fig. 5.4 Idea of frame interpolation. (a) The number of original frames is larger than that of desired frames. (b) The number of original frames is smaller than that of desired frames.

The process of frame generation of the speaking case is illustrated in Fig. 5.5. Firstly, the visemes, namely, the mouth images, of syllables are determined using the frame interpolation technique. Secondly, the visemes of pauses between syllables need be decided. When the duration of a pause is long, it is considered that the person would close his/her mouth; otherwise, the person would keep his/her mouth open and unchanged just if the pause does not exist. Thirdly, the visemes of the first or the last pause should be a closed mouth, because the person does not start speaking during the first pause, and he/she closes his/her mouth after the last syllable. Finally, the determined visemes are integrated into the base images.

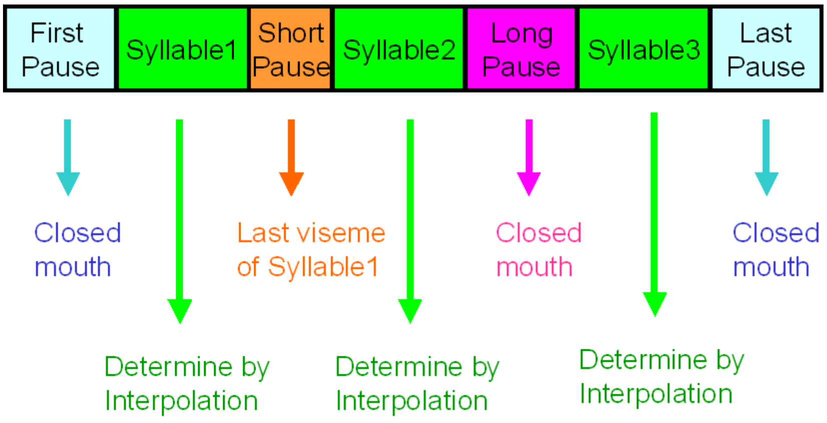


Fig. 5.5 Frame generation of speaking cases.

5.3.2 Singing Case

Due to certain properties of songs, a singing person often has to utter syllables for longer times, especially when he/she is singing a slow song. The frame interpolation technique proposed in Section 5.3.1 is not suitable to determine the visemes of a lengthy syllable because it would make a mouth change its shape in slow motion, which is not natural.

To solve this problem, several facts are noticed. The first is that a mouth always keeps open while singing songs even during long pauses. The second fact is that after the sound of a syllable is uttered, the mouth would hold its shape unchanged and continue uttering the sound. Before the mouth holds its shape, we call that it is in a "mouth-opening" phase. When the mouth begins to hold its shape, we call that it is in a "mouth-holding" phase. Fig. 5.6 shows a diagram of these two phases.

The third noticed 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. In Fig. 5.7, experimental results prove this fact. The duration of the mouth-opening phase using different numbers of beats in a measure is observed. It is shown that when the duration of a syllable is longer, the duration of the mouth-opening phase becomes longer.

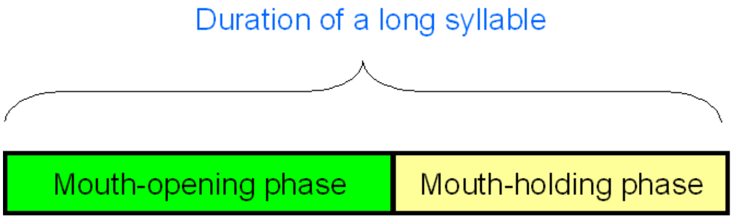
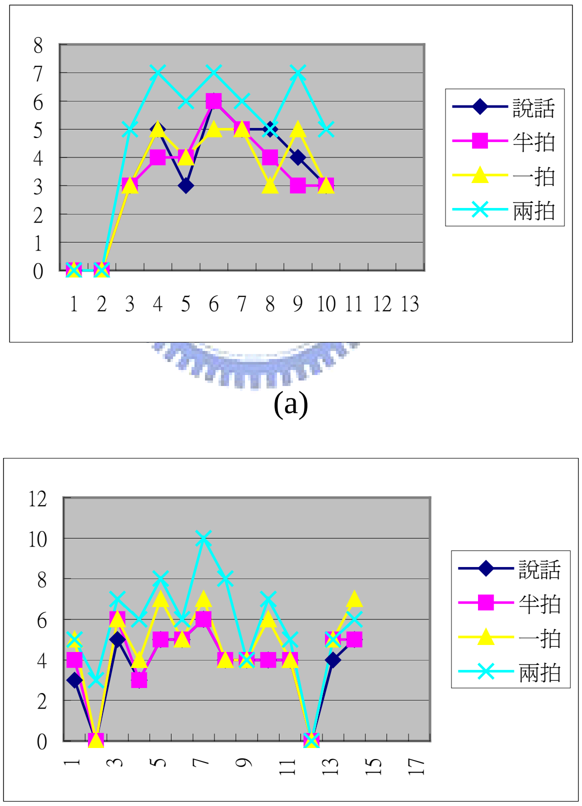


Fig. 5.6 The two phases while singing a long syllable.



(a)

(b)

Fig. 5.7 The duration of the mouth-opening phase of syllables of a same sentence using different beats. (a) The sentence is “紅紅的花開滿了木棉道”. (b) The sentence is “你和我不常聯絡也沒有彼此要求”.

Therefore, the process of frame generation of the speaking case needs to be modified to fit these observed facts. The first modification is that the mouth does not close during pauses. That is, visemes of a pause is the same as the last viseme of the preceding syllable. Orange parts in Fig. 5.8 shows this modification.

The second modification is that the mouth should go through a mouth-opening phase and a mouth-holding phase while singing a long syllable. Suppose 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 $D_{opening}$, and then keep its shape unchanged for a duration of $D_a - D_{opening}$. $D_{opening}$ is defined as follows:

$$D_{opening} = D_d + D_a / D_d$$

Fig. 5.6 shows the entire frame generation process of singing cases.

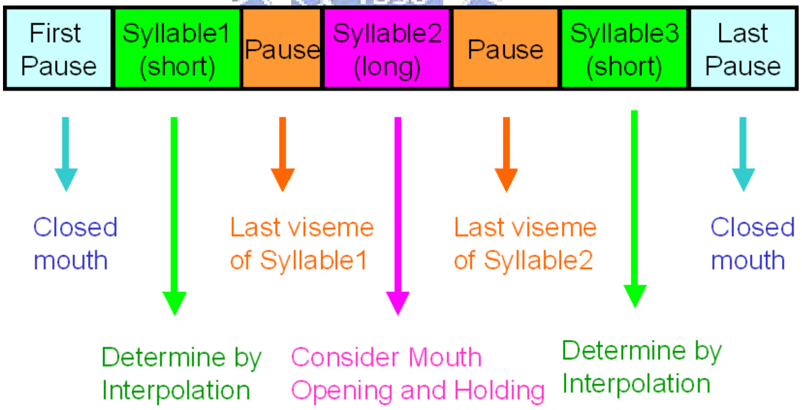


Fig. 5.8 Frame generation of singing cases.

5.4 Smoothing Between Visemes

To create smooth animations, transitions between two successive syllables need to be taken care. When the mouth shape of the last viseme of the preceding syllable is not quite similar to the one of the first viseme of the rear syllable, some transition visemes should be inserted to smooth the articulation.

Fig. 5.8 shows an experimental result concerning the relationship between the distance of two visemes and the number of required transition frames. It is observed that there is no proportional relationship between the distance and the number of transition frames. However, it is noticed that the average number of transition frames is 2. Besides, we notice that while a person is speaking slower, more transition frames can be inserted because the mouth changes its shape slowly. While a person is speaking faster, fewer transition frames needs to be inserted because the mouth changes its shape rapidly.

Therefore, an algorithm for deciding the number of transition frames between two successive syllables is defined as follows:

Algorithm 5. *Deciding the number of transition frames between two syllables*.

Input: a preceding syllable S_1 and a rear syllable $S_{2,}$ a viseme database D, and an input audio A.

Output: the number of required transition frames N.

Steps:

```
Step 1:  Calculate the number of frames $N_{1D}$ of $S_1$ in $D$.
Step 2:  Calculate the number of frames $N_{2D}$ of $S_2$ in $D$.
Step 3:  Calculate the number of frames $N_{1A}$ of $S_1$ in $A$.
Step 4:  Calculate the number of frames $N_{2A}$ of $S_2$ in $A$.
Step 5:  Set $N$ to 3 and stop, if $N_{1A} > N_{1D}$ and $N_{2A} > N_{2D}$.
Step 6:  Set $N$ to 1 and stop, if $N_{1A} < N_{1D}$ and $N_{2A} < N_{2D}$.
Step 7:  Set $N$ to 2.
```

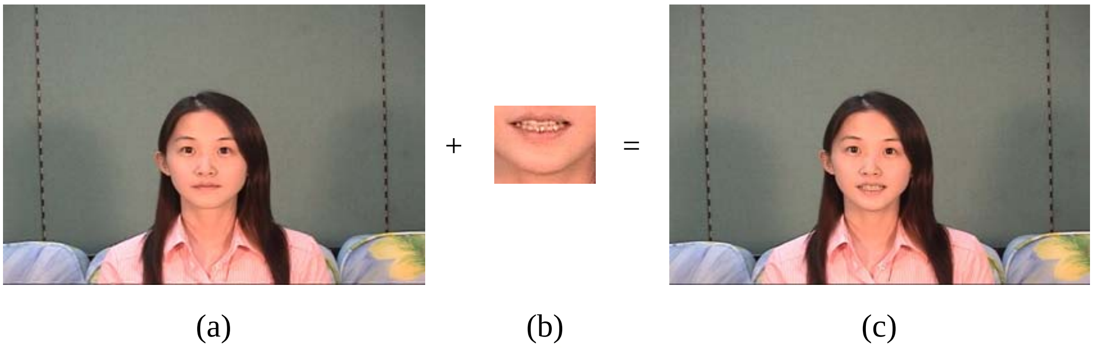
Table 5.1 Relationship between the distance of two visemes and the number of required transition frames.

Transition	Distance of Mouth Shapes	Number of Transition Frames
白→日	18	5
日→依	4	0
依→山	6	1
山→盡	0	2
黃→河	10	5
河→入	12	0
入→海	26	2
海→流	4	1
欲→窮	6	0
窮→千	6	1
千→里	4	3
里→目	12	2
更→上	4	1
上→一	4	1
一→層	4	1
層→樓	0	4

5.5 Integration of Mouth Images and Base Images

After all visemes are determined according to the methods proposed in the previous sections, the mouth image of a viseme needs to be integrated into a base image using the alpha-blending technique, as shown in Fig. 5.9. 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 [6], Lin and Tsai used a fixed rectangle surrounding the mouth to represent the region. The approach is simple to implement;

however, the determination of the size of the rectangle is not an easy job because it affects integrated results severely. A too wide rectangle that overlaps the background such as the walls may cause the background to be “integrated” into the face. A too short rectangle may cause the jaw to “drop down” while opening the mouth because only part of the jaw is moving.



(a) (b) (c)

Fig. 5.9 Integration of a base image and a mouth image. (a) A base image. (b) A mouth image. (c) The integrated image.

A method is proposed to determine the region of a mouth image. 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. Then the method goes as follows.

Firstly, since the position and size of the base region are known already using the methods proposed in Chapter 4, we can determine the skin color by averaging the colors in the base region. Secondly, skin regions of I_1 and I_2 are determined 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.

Besides using $S_{\text{intersect}}$ as the mouth region, a trapezoid inside it can also be used as another choice of the mouth region. Fig. 5.10 shows examples of these two kinds of mouth regions.

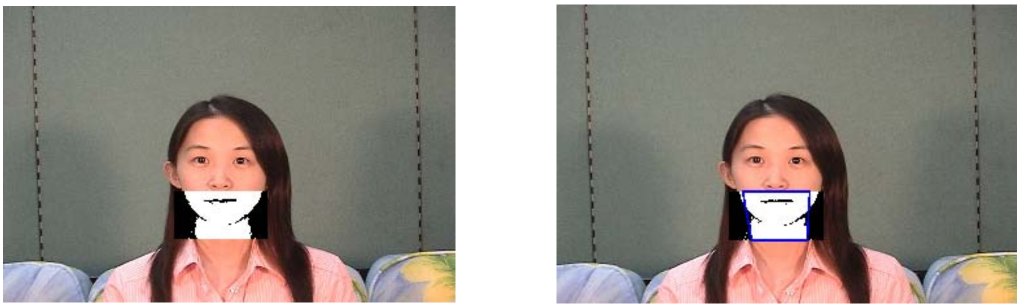


Fig. 5.10 Example of mouth regions found using the proposed method. (a) The intersection region of skin parts. (b) A trapezoid inside the intersection region.

5.6 Experimental Results

In this section, some experimental results of the proposed methods are shown. In Fig. 5.11, the model is speaking a sentence. And in Fig. 5.12, the model is singing a song.

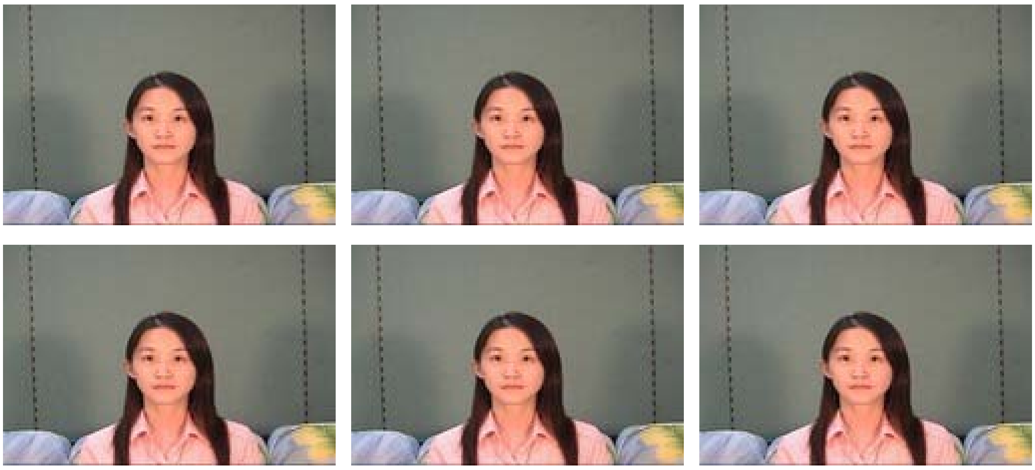


Fig. 5.11 Result of frame generation of a speaking case. The person is speaking the sentence “夕陽”.

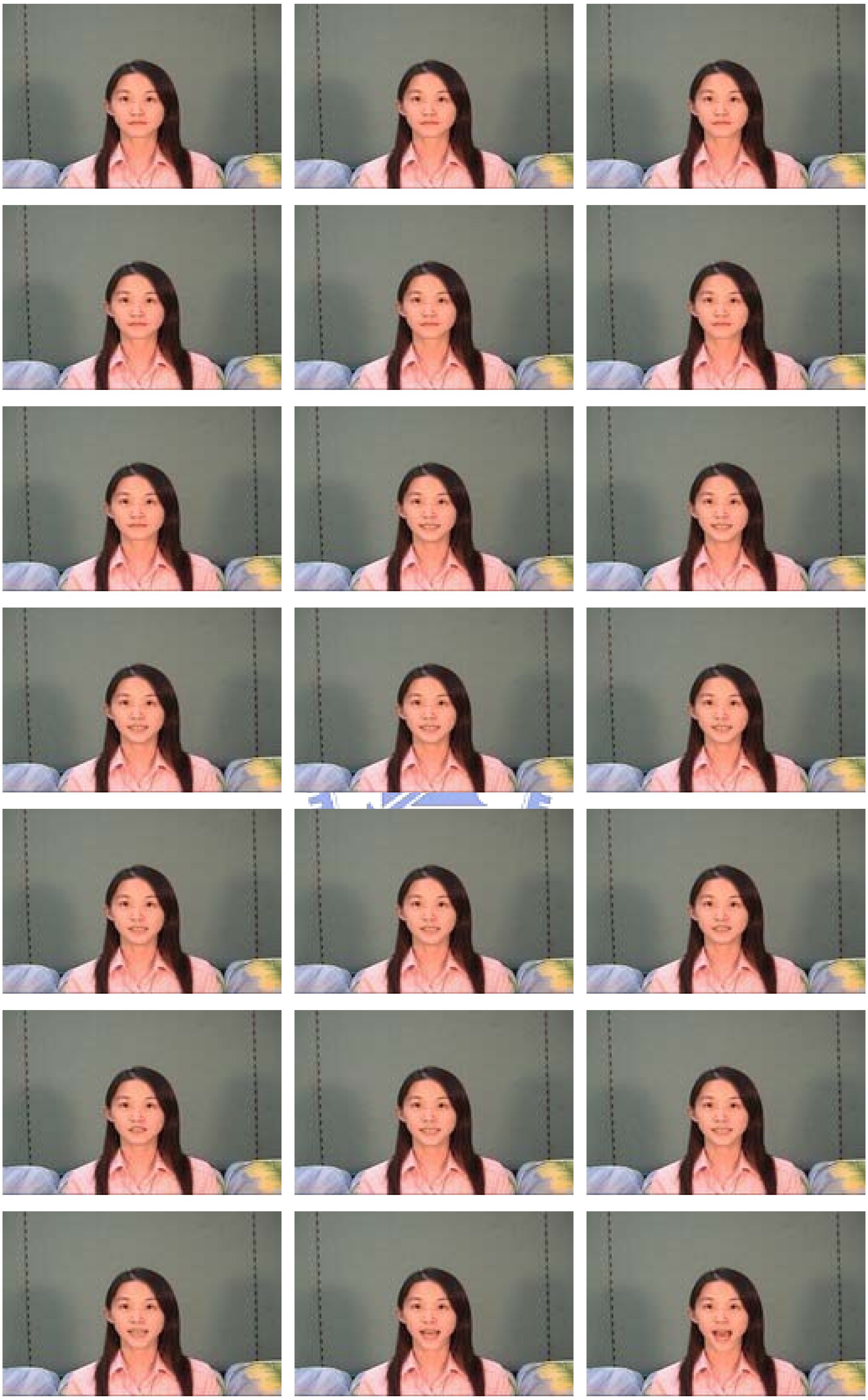


Fig. 5.11 Result of frame generation of a speaking case. The person is speaking the sentence “夕陽”. (Continued)

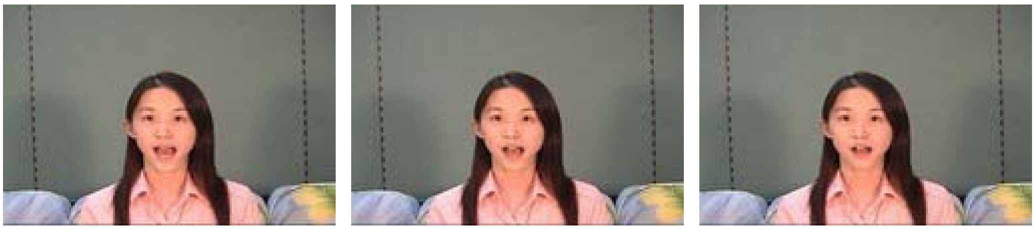


Fig. 5.11 Result of frame generation of a speaking case. The person is speaking the sentence “夕陽”. (Continued)

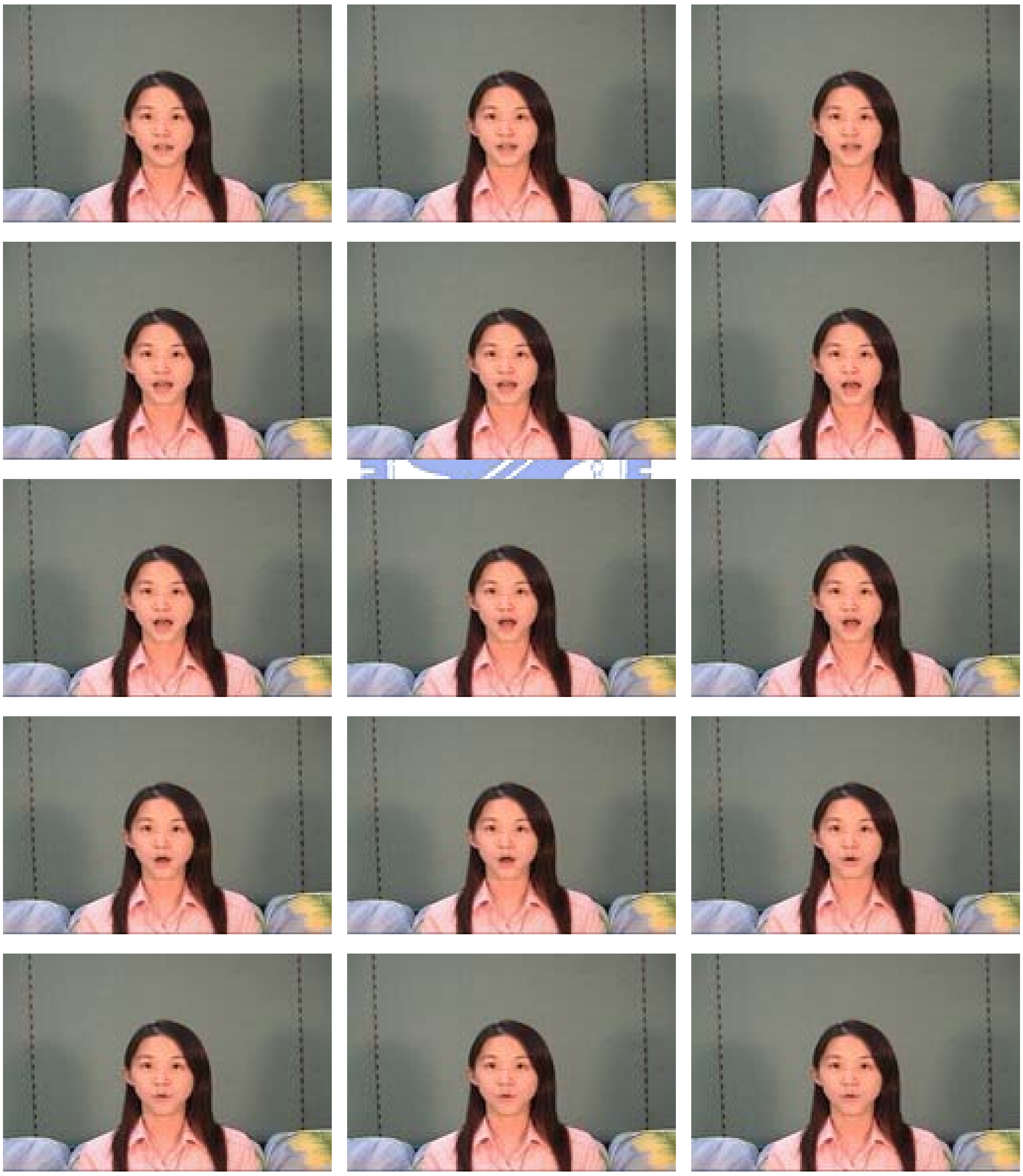


Fig. 5.12 Result of frame generation of a singing case. The person is speaking the sentence “如果雲知道”. The frames shown are part of “知道”.

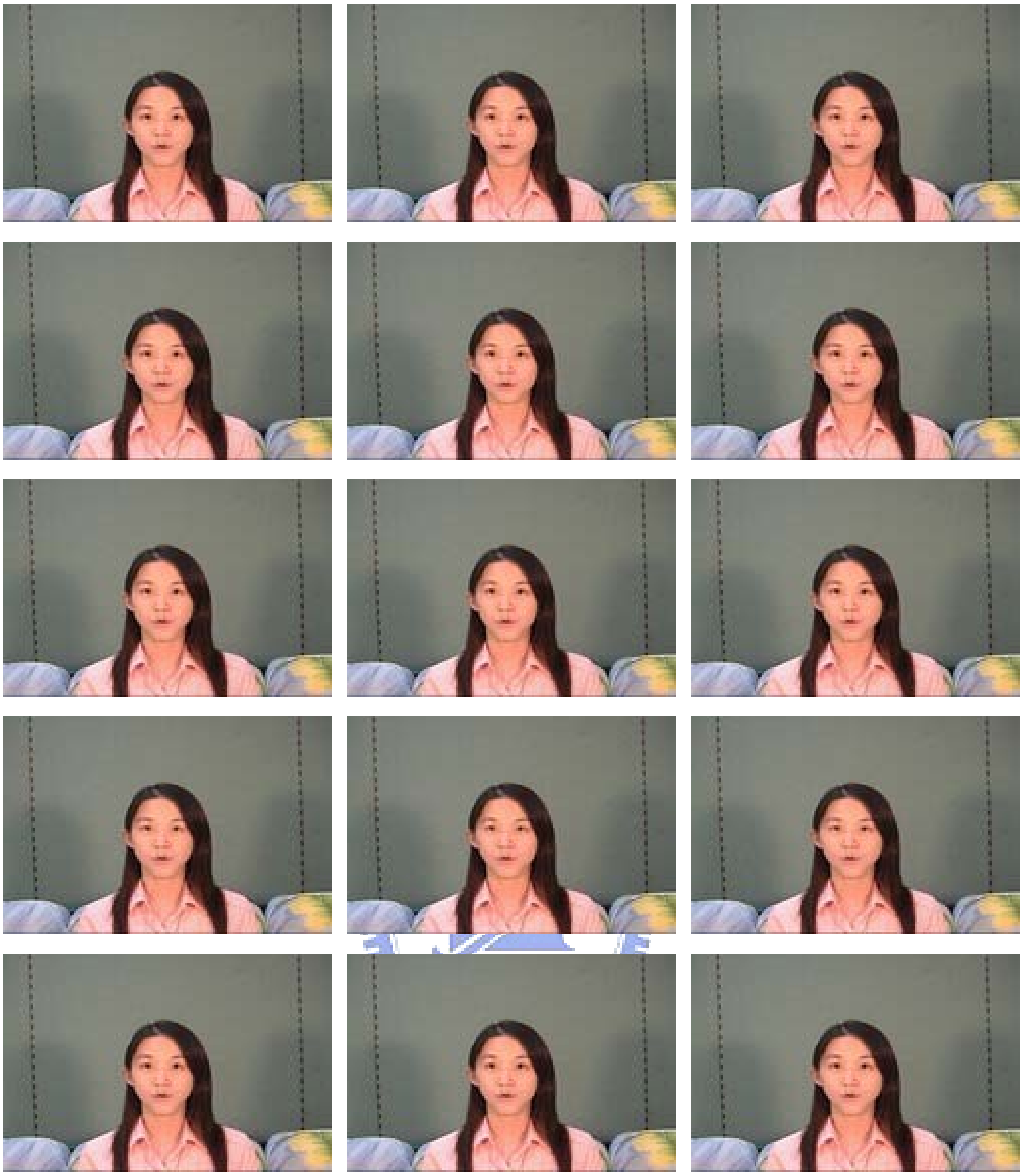


Fig. 5.12 Result of frame generation of a singing case. The person is speaking the sentence “如果雲知道”. The frames shown are part of “知道”. (Continued)

5.7 Summary and Discussions

In this chapter, the concentration has been put on improving the quality of generated animations. The first issue was synchronization between a speech and image frames because people can notice the asynchronous problem easily. The proposed method reduced the synchronization error down to be shorter in time than the period of a frame. To make the animations natural, the behaviors of real talking persons and singing persons were discussed so that generated virtual faces can act in the same way as real human beings. The articulation effect of transitions was also noticed and solved. Finally, mouth regions found by the proposed method were proper to be integrated into base images.

Chapter 6
Examples of Applications

6.1 Introduction

Virtual talking faces can be applied to many areas. For example, they can be used as agents or assistants to help people do their jobs. They can also be used as tutors to help students study.

In this chapter, some examples of applications are described. Section 6.2 shows how virtual announcers that are able to report news are created. Section 6.3 presents virtual singers that can sing songs. In Section 6.4, virtual talking faces are integrated into emails, so that people can watch their friends reading the contents of received emails. In Section 6.5, some other possible applications are listed.

6.2 Virtual Announcers

6.2.1 Description

Virtual announcers are virtual talking faces that can report news. As shown in Fig. 6.1, real news reporters appear on television screens with their faces and part of their bodies seen. Since real news reporters may sometimes get sick or be occupied by other tasks, virtual announcers can take the place of them who shall record news releases in advance. Moreover, they can even replace real news reporters as the techniques are exquisitely applied and they exhibit very realistic appearances.

Fig. 6.1 Examples of real news reporters.

6.2.2 Process of Creation

The process of creating a virtual announcer is shown in Fig. 6.2. Firstly, the video of a speaking person needs to be recorded. Secondly, the feature learning process extracts all required features from the video. These two jobs follow the processes proposed in Chapters 2 through 4. Then, any audio of news can be sent to the animation generation process to create animations of virtual announcers. It is noticeable that the frame generation method for the speaking case proposed in Section 5.3.1 is utilized at this stage.

An example of a virtual announcer is shown in Fig. 6.3. The background may be changed dynamically to fit the news that is being reported.

6.3 Virtual Singers

6.3.1 Description

Similar to virtual announcers, virtual singers make use of virtual faces that are able to sing songs. They can be used for entertainment.

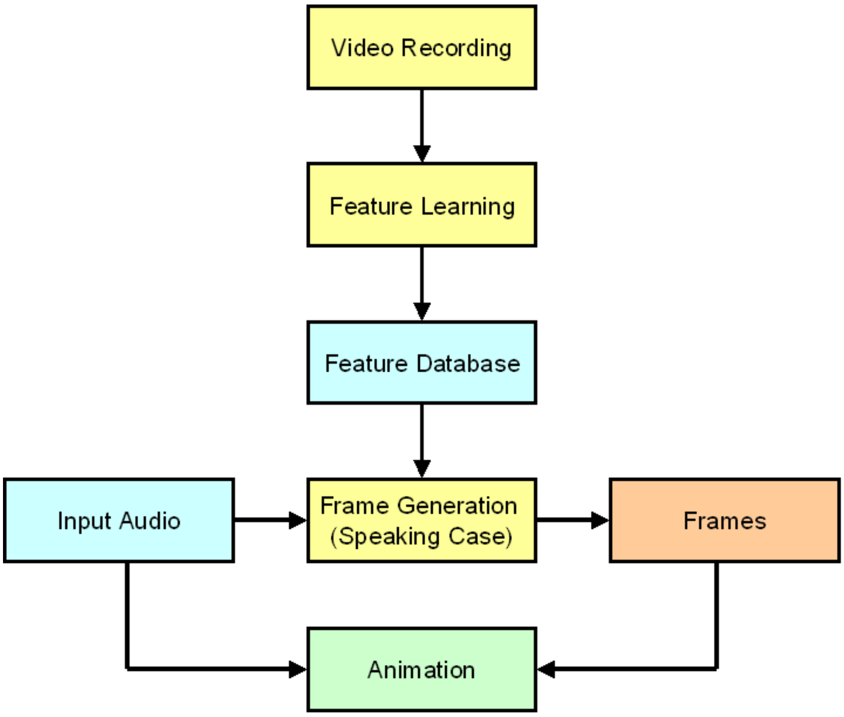


Fig. 6.2 Block diagram of creation of virtual announcers.

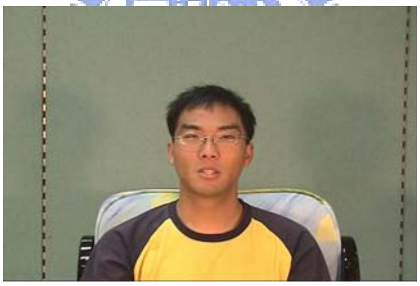


Fig. 6.3 Example of a virtual announcer.

6.3.2 Process of Creation

The process of creating a virtual singer is shown in Fig. 6.4. The process is mostly the same as the one of creating a virtual announcer. The only difference is that that the frame generation method for the singing case proposed in Section 5.3.2, instead of that of the speaking case, is utilized.

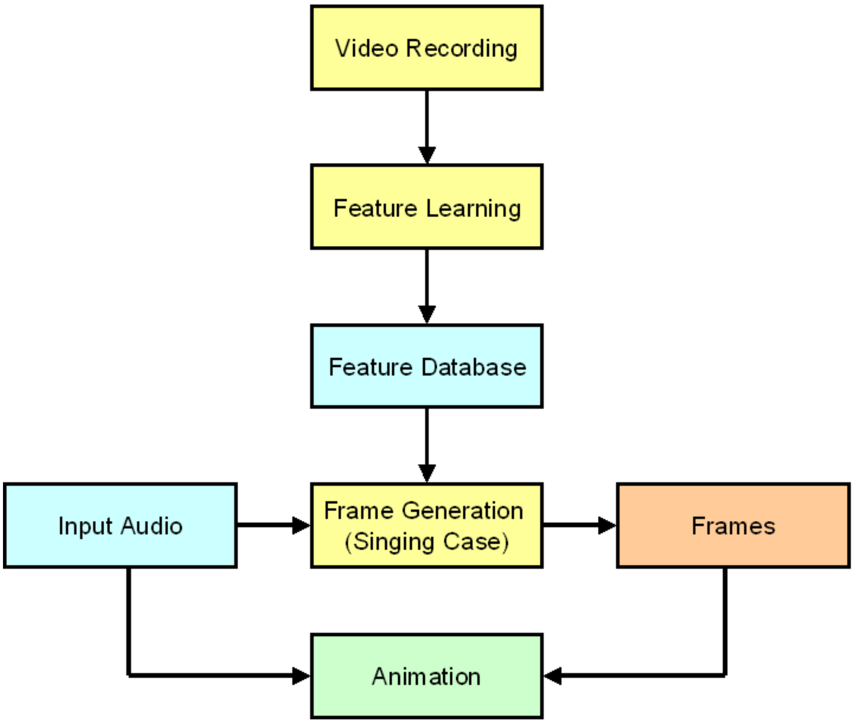


Fig. 6.4 Block diagram of creation of virtual singers.

6.4 Emails with Virtual Talking Faces

6.4.1 Description

The use of emails is a very common way to transmit messages nowadays. People send and receive emails almost everyday. When a person receives an email, he/she needs to read the content of the email to know what the sender wants to express. A virtual talking face can be integrated into an email to enable the receiver to understand the content of the email by watching and listening to the animation of the sender’s face without reading the content.

Fig. 6.5 shows comparisons between normal emails and emails with embedded virtual talking faces. Fig. 6.5(a) shows that a text email is sent through the Internet and then received. The receiver needs to read the text message. Fig. 6.5(b) shows the process of sending and receiving an email with virtual talking faces. An email text and its corresponding speech read by the sender are sent through the Internet. The receiver can generate a virtual talking face that reads the text and then watch it.

Fig. 6.5(c) shows a way to produce a similar result of (b). The sender records a video while he/she is reading the text, and then the video along with the text is sent through the Internet. The receiver can watch the video directly. However, this method has some disadvantages. First, the sender needs to record the scene while he/she is reading, that means a camera is required. Second, the video size is often much larger than that of the speech. Sending large videos through the Internet is slow, and the receiver gets annoyed while receiving big emails.

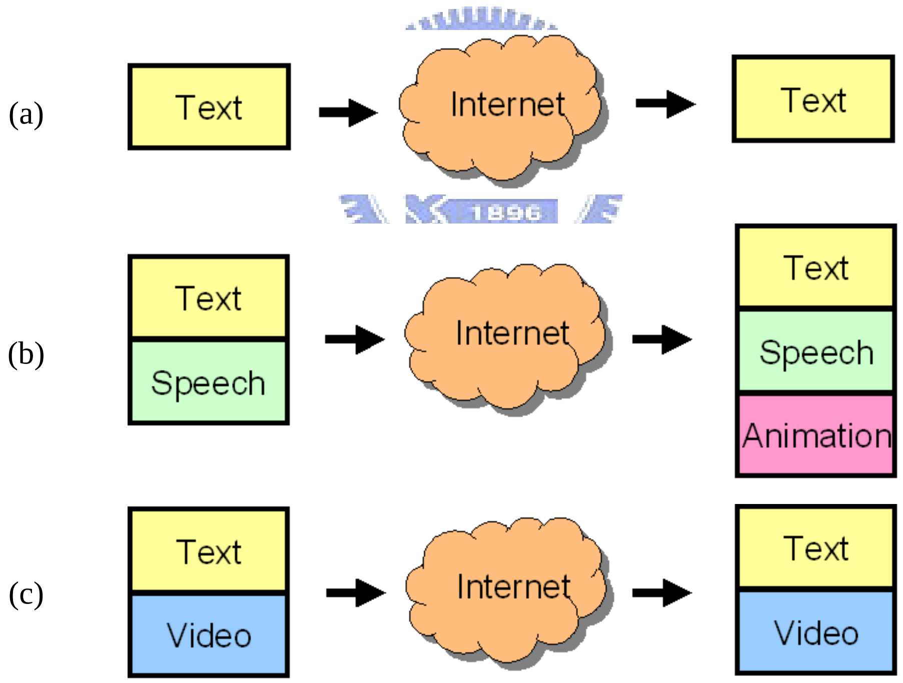


Fig. 6.5 Comparisons among three kinds of emails.

6.4.2 Process of Sending Emails

As shown in Fig. 6.5(b), the speech needs to be sent along with the email content.

The speech can be wrapped as an attachment file of an email. Some extra information can be added to the attachment file, such as the name of the sender. Fig. 6.6 shows the structure of the attachment file.

Fig. 6.7 shows the entire process of sending emails with virtual talking faces. Firstly, the sender writes a text and then reads it. The speech is recorded as an audio file. Then the information of the name of the sender, the size of the text, and the speech are combined together with the text and the speech to form an attachment file. Finally, the attachment file is sent through the Internet.

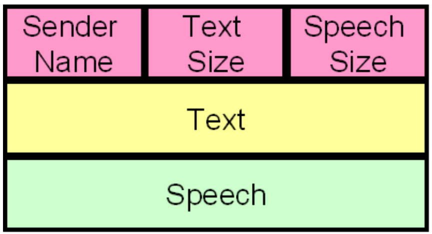


Fig. 6.6 Structure of an attachment file.

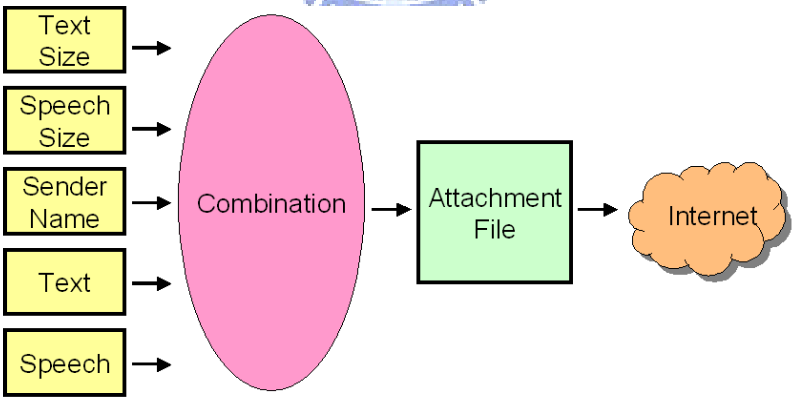


Fig. 6.7 Process of sending emails with a virtual talking face.

6.4.3 Process of Receiving Emails

After receiving an email mentioned in the above section, the receiver can construct an animation of the sender’s face. The process is illustrated in Fig. 6.8. Firstly the text size and the speech size are extracted to help segment the text and the speech out. The sender name is used to select the face to generate. Then, the text and the speech are sent to the animation generation process as inputs to produce an animation.

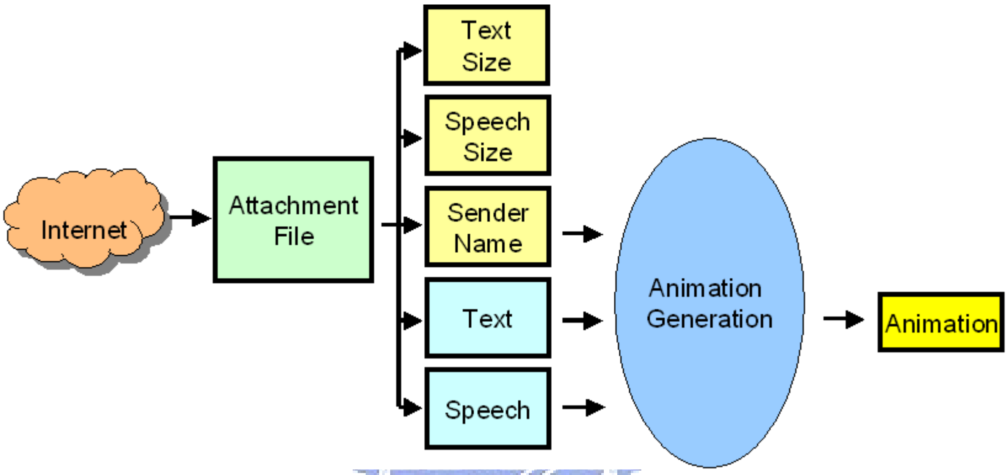


Fig. 6.8 Process of receiving an email with a virtual talking face.

6.5 Other Applications

Besides the applications mentioned in the previous sections of this chapter, there are still many possible applications of virtual talking faces. Some of them are listed as follows.

1. **Virtual tutors**: Virtual talking faces can be used to help students study. It is interesting and useful for the students to have tutors teaching rather than to study by themselves.
2. **Virtual guides**: Virtual talking faces can be used in libraries and museums to guide visitors. Visitors can roam around in the buildings and listen to the

illustration by a virtual guide.

3. Virtual babysitters: Virtual talking faces can read stories for children as babysitters.
4. Software agents: Virtual talking faces can be used in the software to help users operate the software.
5. Videoconferences: The amount of data required for virtual talking faces to be transmitted through networks is much less than that for transmitting image frames.

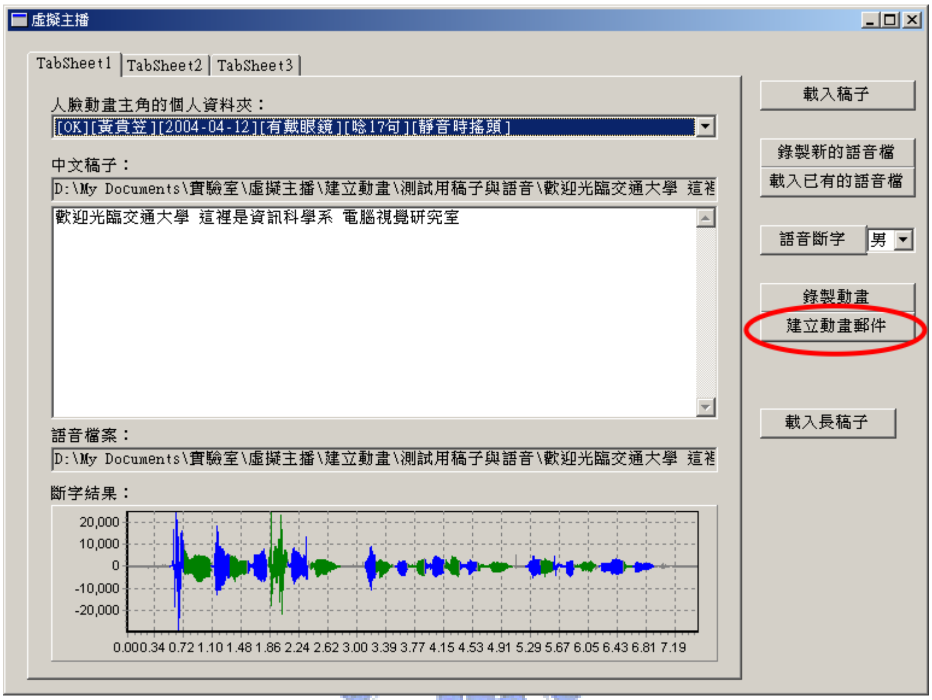
6.6 Experimental Results

In this section, some experimental results of creating emails with virtual talking faces are shown. The experimental results of virtual announcers and singers are shown in Section 5.6 in Chapter 5 already.

Fig. 6.9 shows the program interface of animation generation. Firstly, the email sender inputs a text and a speech, and then presses the button marked in red to create an attachment file, as shown in Fig. 6.10. The red circle indicates the attachment file.

As shown in Fig. 6.11, after receiving the email, the sender also receives the attachment. It is noticed that the size of the attachment file is only 120 KB, which is very small.

The receiver then opens the attachment file, and the file relationship will cause the program to start the proposed animation generation process, as shown in Fig. 6.12 and 6.13. The final animation generated is shown in Fig. 6.14(a). Fig. 6.14(b) shows that the size of the animation is 3.67 MB, which is much larger than that of the attachment file. Therefore, it is obvious that transmitting the virtual talking faces using the method proposed in Section 6.4 is preferred to transmitting the videos directly.



人臉動畫主角的個人資料夾：[OK][黃貴笠][2004-04-12][有戴眼鏡][唸17句][靜音時搖頭]

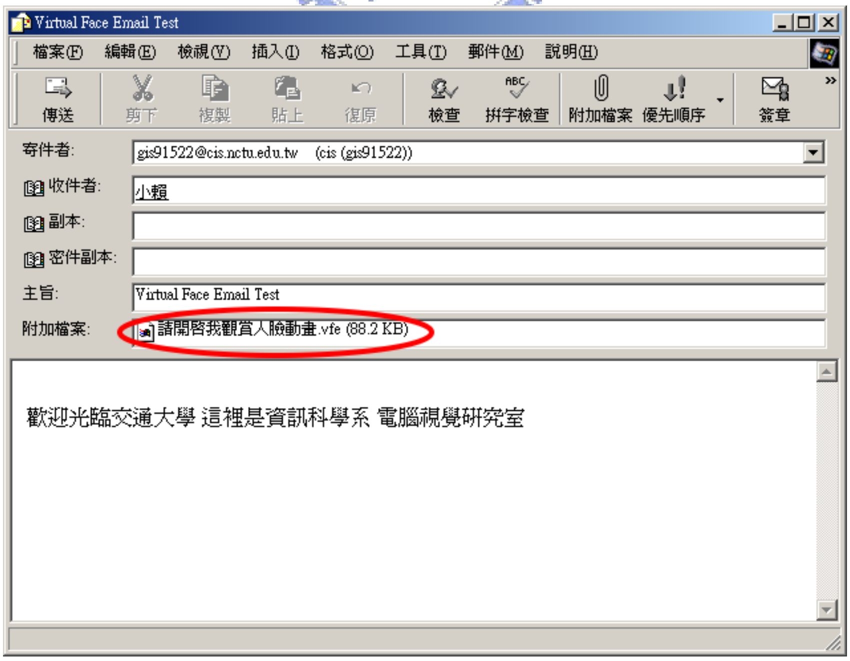
中文稿子：D:\My Documents\實驗室\虛擬主播\建立動畫\測試用稿子與語音\歡迎光臨交通大學 這裡

歡迎光臨交通大學 這裡是資訊科學系 電腦視覺研究室

語音檔案：D:\My Documents\實驗室\虛擬主播\建立動畫\測試用稿子與語音\歡迎光臨交通大學 這裡

斷字結果：

Fig. 6.9 The program interface of animation generation.



寄件者: gis91522@cis.nctu.edu.tw (cis (gis91522))

收件者: 小賴

副本:

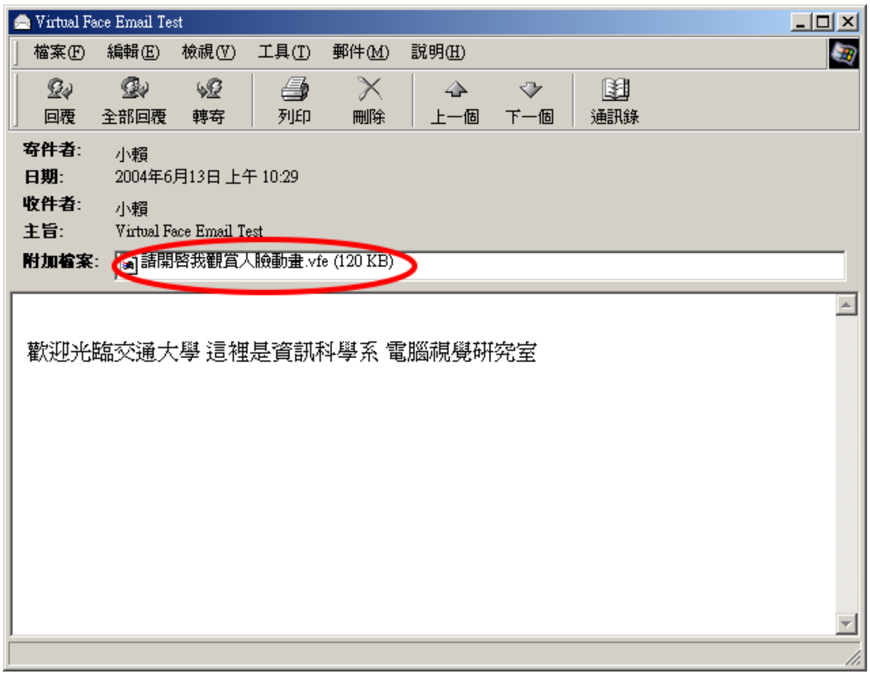
密件副本:

主旨: Virtual Face Email Test

附加檔案: 請開啓我觀賞人臉動畫.vfe (88.2 KB)

歡迎光臨交通大學 這裡是資訊科學系 電腦視覺研究室

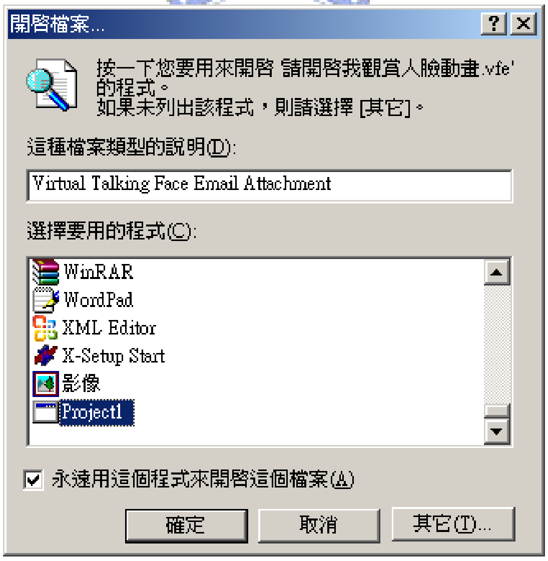
Fig. 6.10 The email attachment is created.



寄件者: 小賴
日期: 2004年6月13日 上午 10:29
收件者: 小賴
主旨: Virtual Face Email Test
附加檔案: 請開啓我觀賞人臉動畫.vfe (120 KB)

歡迎光臨交通大學 這裡是資訊科學系 電腦視覺研究室

Fig. 6.11 The attachment is received along with the email text.



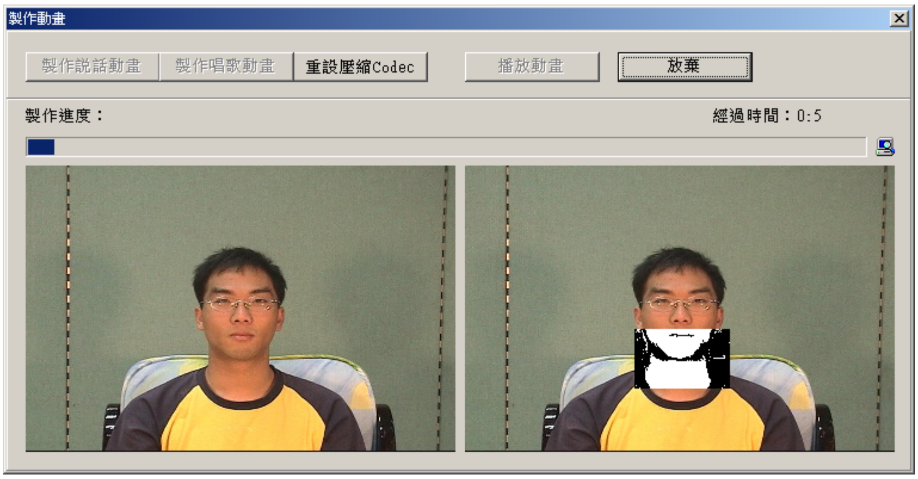
按一下您要用來開啓 '請開啓我觀賞人臉動畫.vfe' 的程式。
如果未列出該程式，則請選擇 [其它]。

這種檔案類型的說明(D): Virtual Talking Face Email Attachment

選擇要用的程式(C): Project1

☑ 永遠用這個程式來開啓這個檔案(A)

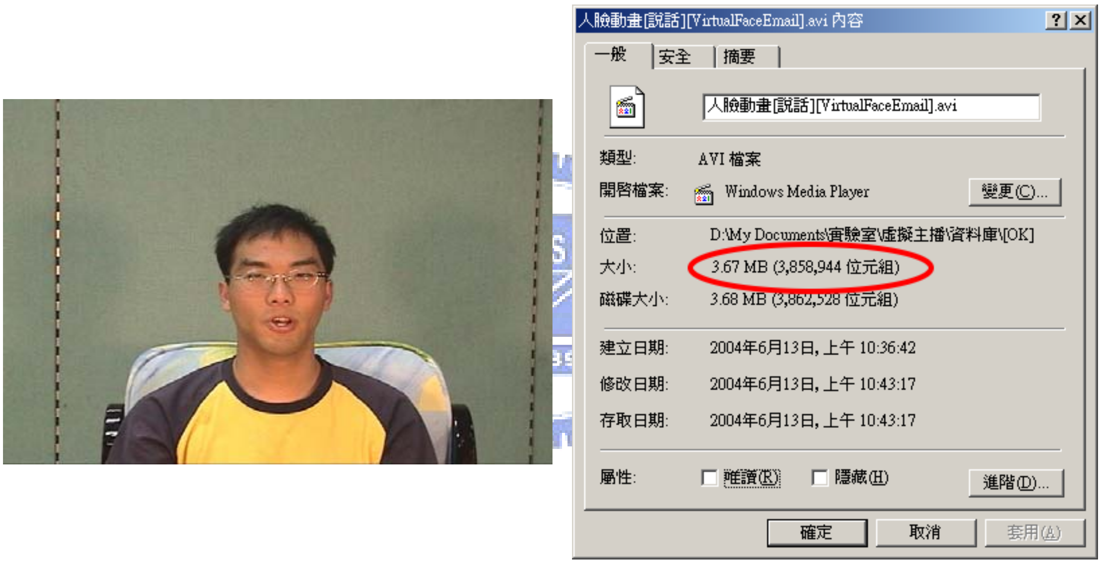
Fig. 6.12 Setting of the file relationship between the attachment file and the program.



製作進度：

經過時間：0:5

Fig. 6.13. The animation generation process is generating the animation of a person.



人臉動畫[說話][VirtualFaceEmail].avi

類型: AVI 檔案

開啓檔案: Windows Media Player

位置: D:\My Documents\實驗室\虛擬主播\資料庫\[OK]

大小: 3.67 MB (3,858,944 位元組)

磁碟大小: 3.68 MB (3,862,528 位元組)

建立日期: 2004年6月13日, 上午 10:36:42

修改日期: 2004年6月13日, 上午 10:43:17

存取日期: 2004年6月13日, 上午 10:43:17

屬性: □ 唯讀(R) □ 隱藏(H)

(a) (b)

Fig. 6.14 The generated animation. (a) A frame of the animation. (b) The size of the generated animation.

6.7 Summary and Discussions

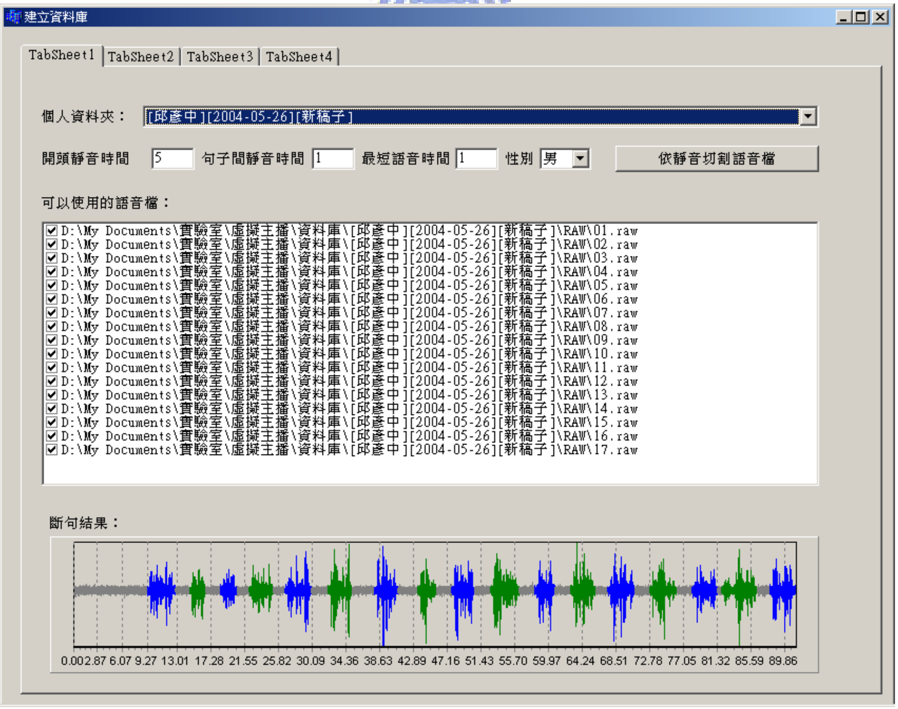
In this chapter, three examples of virtual talking faces are described and implemented. The virtual announcers can be used to report news while the virtual singers can sing songs. A system for transmitting virtual talking faces by emails is also proposed. It is very interesting and useful to watch friends read their mails. And the amount of transmitted data is much smaller than that of transmitting similar videos.

Chapter 7
Experimental Results and Discussions

7.1 Experimental Results

In this study, a system for automatic feature learning and animation creation is constructed. Some screenshots of the system are shown in this section.

Firstly, a video that contains a speaking face was recorded, and then audio data and image frames were extracted from it. In Fig. 7.1, the extracted audio data were segmented into seventeen sound parts using the sentence segmentation algorithm proposed in Chapter 3. Blue and green parts represent odd and even sound parts, respectively.



個人資料夾：[邱彥中][2004-05-26][新稿子]

開頭靜音時間 5　句子間靜音時間 1　最短語音時間 1　性別 男

可以使用的語音檔：

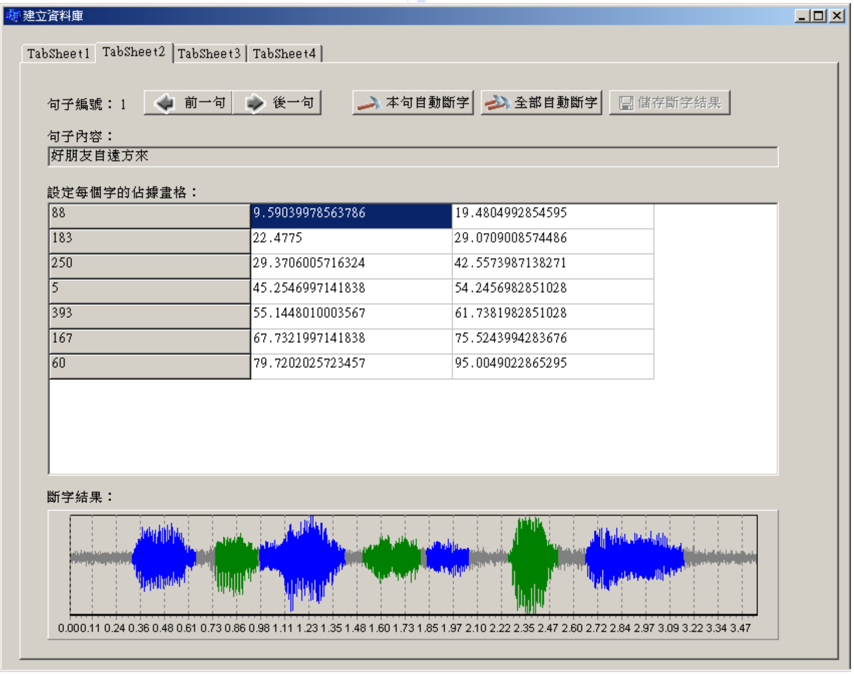
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\01.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\02.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\03.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\04.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\05.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\06.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\07.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\08.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\09.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\10.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\11.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\12.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\13.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\14.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\15.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\16.raw
☑ D:\My Documents\實驗室\虛擬主播\資料庫\[邱彥中][2004-05-26][新稿子]\RAW\17.raw

斷句結果：

Fig. 7.1 The audio data are segmented into seventeen sound parts.

Next, the timing information of syllables of these sound parts was learned using the syllable alignment technique, as shown in Fig. 7.2. Blue and green parts represent odd and even syllables, respectively. Thirdly, the base region of the first frame of the video was determined. Fig. 7.3 shows the base region with a blue rectangle. Fig. 7.4 shows the final step of the feature learning process. The position and the rotated angle of the base region for every frame were learned using the methods proposed in Chapter 4. The black cross on the right face in Fig. 7.4 indicates the center of the base region.

After the feature learning process was completed, the animation generation process was started. Firstly, the syllable alignment technique was applied to an input speech to get the timing information of syllables in the speech. In Fig. 7.5, a speech containing a Mandarin sentence “熟能生巧” was used as an input. The timing information of syllables in it was learned and displayed in blue and green colors, which represent odd and even syllables, respectively.



句子編號：1

句子內容：
好朋友自遠方來

設定每個字的佔據畫格：

88	9.59039978563786	19.4804992854595
183	22.4775	29.0709008574486
250	29.3706005716324	42.5573987138271
5	45.2546997141838	54.2456982851028
393	55.1448010003567	61.7381982851028
167	67.7321997141838	75.5243994283676
60	79.7202025723457	95.0049022865295

Fig. 7.2 The result of syllable alignment of the sentence “好朋友自遠方來”.

After the timing information of the input speech is known, virtual talking face animations can be generated. Using the methods proposed in Chapter 5, we could generate proper frames. Fig. 7.6 shows the intermediate result of the process. The right face shows the region that was to be "pasted" onto the base image, and the left face shows the result of integration. Fig. 7.7 shows the final created animation in frames. The face in the frames is speaking the Chinese sentence "熟能生巧".

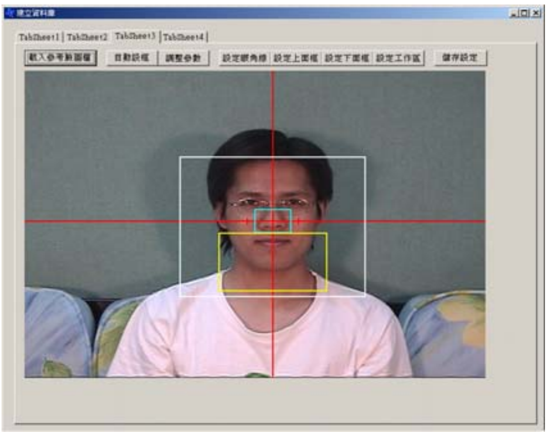
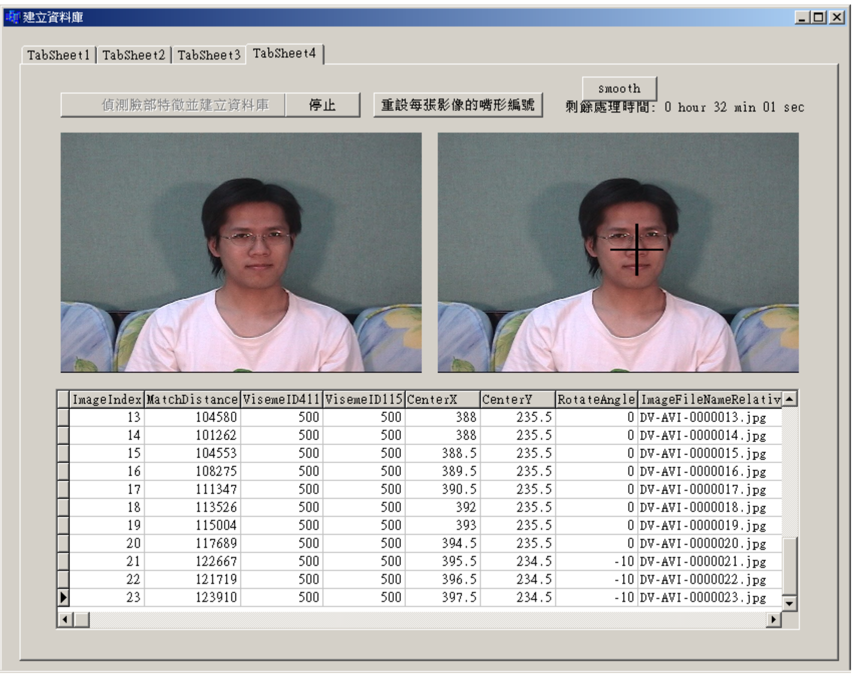


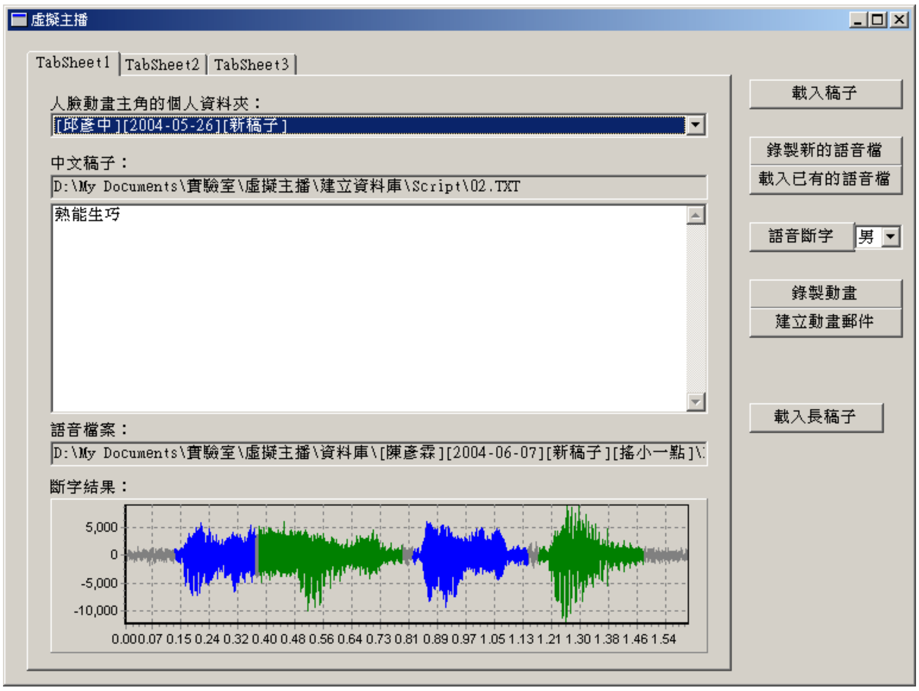
Fig. 7.3 The base region determined is displayed with a blue rectangle.



剩餘處理時間: 0 hour 32 min 01 sec

ImageIndex	MatchDistance	VisemeID411	VisemeID115	CenterX	CenterY	RotateAngle	ImageFileNameRelativ
13	104580	500	500	388	235.5	0	DV-AVI-0000013.jpg
14	101262	500	500	388	235.5	0	DV-AVI-0000014.jpg
15	104553	500	500	388.5	235.5	0	DV-AVI-0000015.jpg
16	108275	500	500	389.5	235.5	0	DV-AVI-0000016.jpg
17	111347	500	500	390.5	235.5	0	DV-AVI-0000017.jpg
18	113526	500	500	392	235.5	0	DV-AVI-0000018.jpg
19	115004	500	500	393	235.5	0	DV-AVI-0000019.jpg
20	117689	500	500	394.5	235.5	0	DV-AVI-0000020.jpg
21	122667	500	500	395.5	234.5	-10	DV-AVI-0000021.jpg
22	121719	500	500	396.5	234.5	-10	DV-AVI-0000022.jpg
23	123910	500	500	397.5	234.5	-10	DV-AVI-0000023.jpg

Fig. 7.4 Learning of the position of the base region for every frame.



虛擬主播

人臉動畫主角的個人資料夾：[邱彥中][2004-05-26][新稿子]

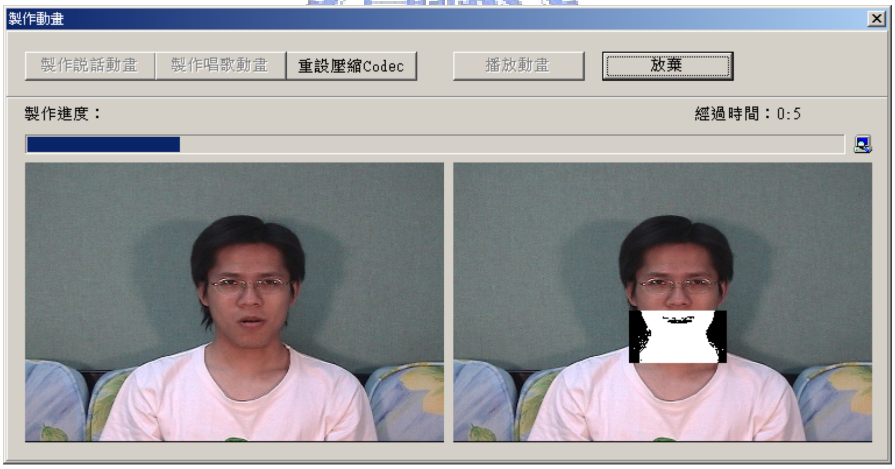
中文稿子：D:\My Documents\實驗室\虛擬主播\建立資料庫\Script\02.TXT

熟能生巧

語音檔案：D:\My Documents\實驗室\虛擬主播\資料庫\[陳彥霖][2004-06-07][新稿子][搖小一點]\

斷字結果：

Fig. 7.5 The result of syllable alignment of the input speech “熟能生巧”.



製作動畫

製作進度：

經過時間：0:5

Fig. 7.6 The middle result of the frame generation process.

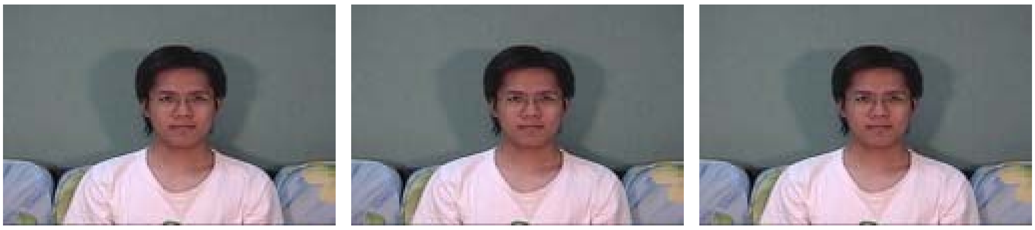


Fig. 7.7 The result of the animation generation process.

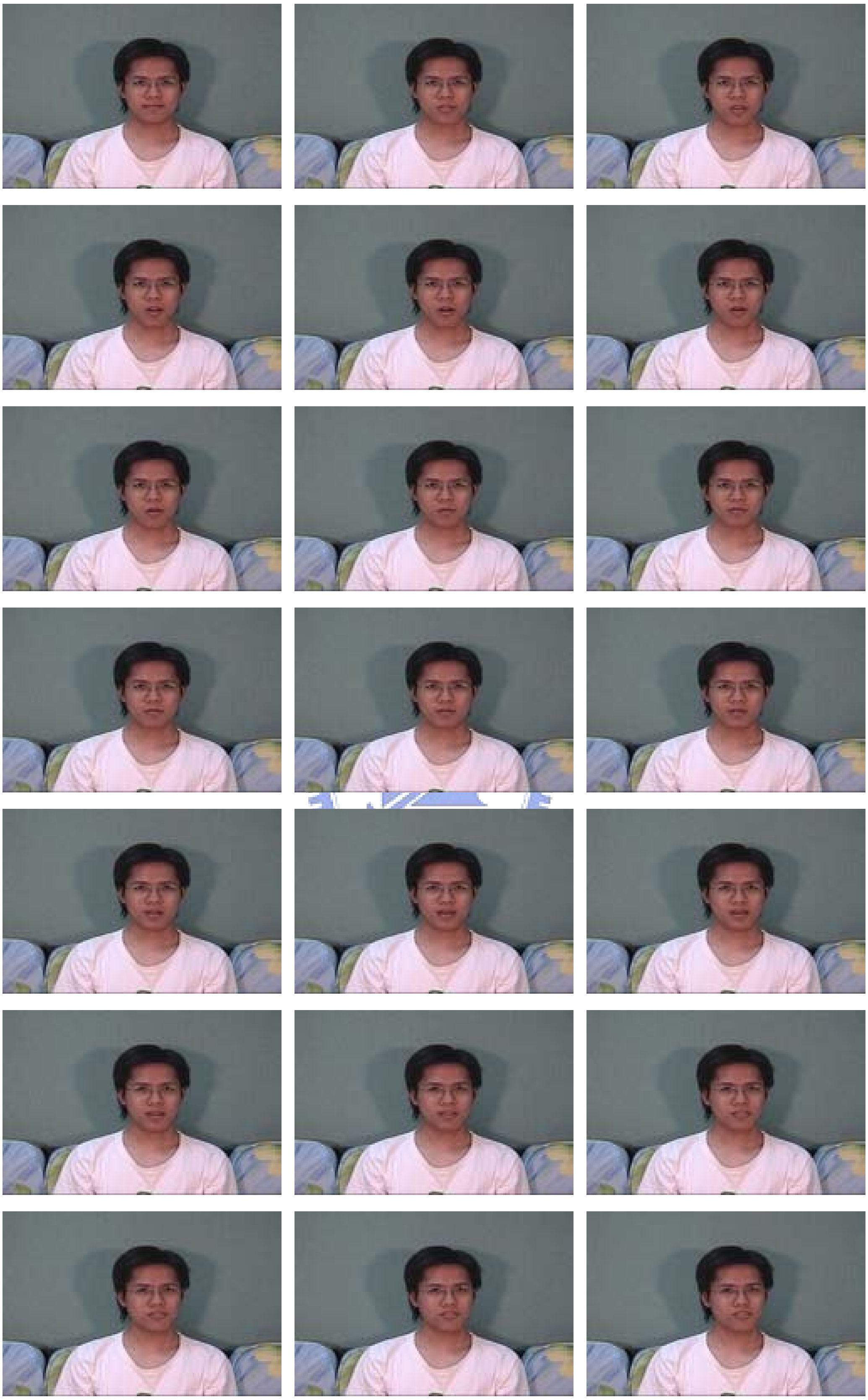


Fig. 7.7 The result of the animation generation process. (Continued)

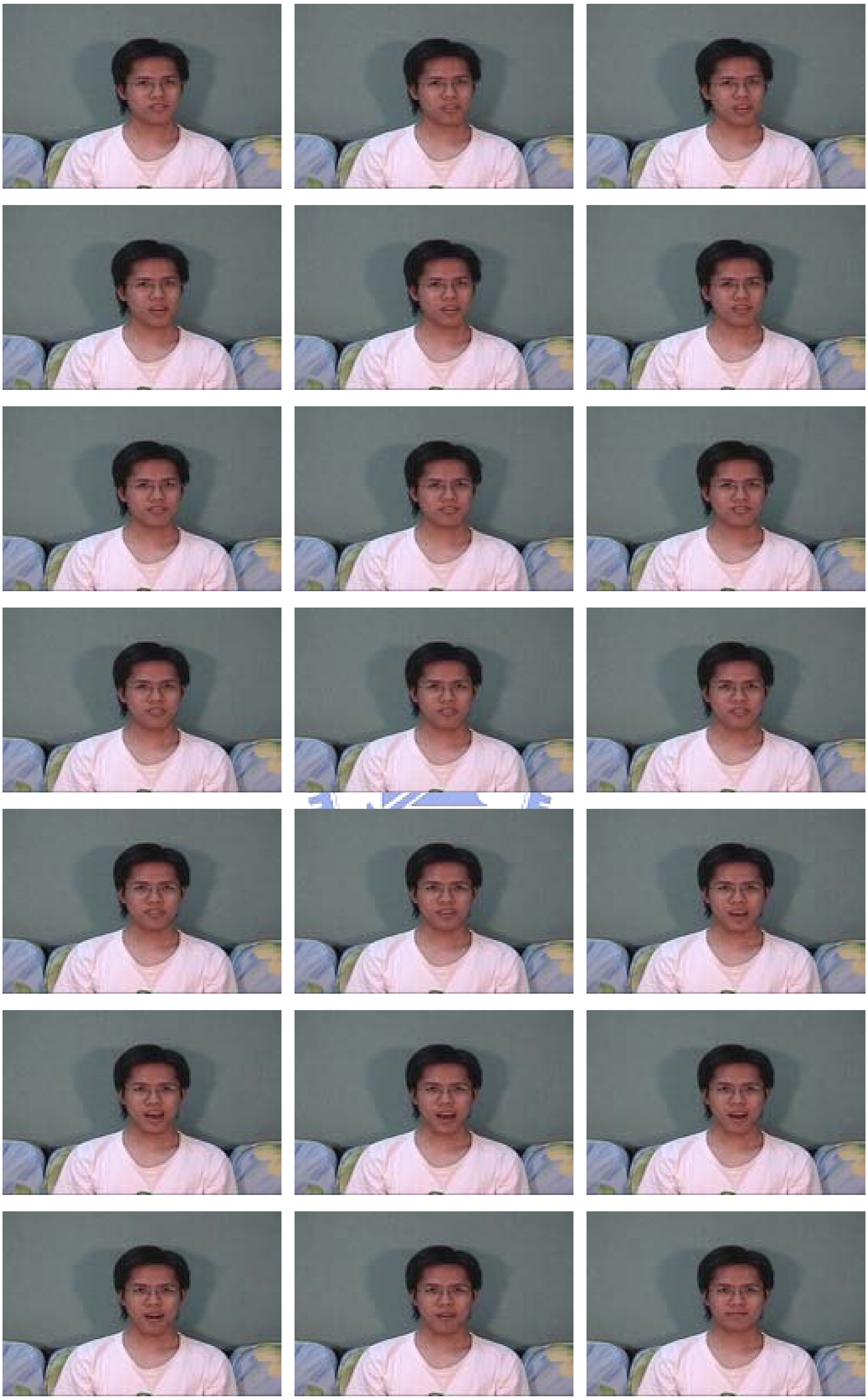


Fig. 7.7 The result of the animation generation process. (Continued)

7.2 Discussions

After presenting the experimental results, we would like to discuss a number of issues in concern as follows.

The first issue is the video recording process. Since a transcript that contains all classes of Mandarin syllables is designed to consist of seventeen short and meaningful sentences, the model can read them easily. The model can also shake his/her head slightly during the process. The process takes about two minutes to complete, which is quite short and not annoying.

The second issue is the feature learning process. To learn the information of audio features, the audio data of the recorded video are segmented into sentences first. The proposed sentence segmentation algorithm is effective both in quiet environments and in environments with constant noise - like that caused by fans. The result of syllable alignment may not be completely correct; however, it is still acceptable.

To learn the information of facial features, the position and rotated angle of the base region in every frame are detected. The information learned is correct even when the face is shaking slightly. Besides, the process is automatic; therefore, we can avoid finding the positions manually, which is a tedious work. Normally, the process takes about half an hour to complete, which is a little lengthy but acceptable.

Finally, we discuss the animation generation process. In this process, proper frames are generated. The experimental results show that the generated talking faces are natural. The faces also shake naturally due to the arrangement of base images. It is shown that 2D images are suitable for creating realistic talking faces with slight head shaking actions.

Chapter 8
Conclusions and Suggestions for Future Works

8.1 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. Techniques of image processing and speech recognition are utilized. Methods are proposed to automate the learning process and improve the quality of the generated animations.

The system is composed of three processes: video recording, feature learning, and animation generation. The video recording process is designed to be short, easy, and not annoying. A transcript that contains all classes of Mandarin syllables has been proposed, and a model can read the sentences on it, instead of reading the syllables one by one separately. The sentences were designed to be short and meaningful, so that the model can read them without difficulties. Due to the function of the feature learning process, 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 sentence segmentation algorithm proposed in the learning of audio features is simple but robust in both quiet environments and environments with constant noise. The generation process of base image sequences was designed to exhibit natural head shaking actions. The learning process of facial features was designed to be able to handle shaking faces.

In the animation generation process, several methods were proposed to improve the quality of generated animations. A method was proposed to reduce the synchronization error between a speech and image frames down to be shorter in time than the period of a frame. The number of required transition frames between successive syllables is analyzed to smooth the transitions. The behaviors of a talking person and a singing person are also analyzed, and a method of frame generation that is proper to create both talking and singing faces was proposed. Applications that utilize the proposed techniques such as virtual announcers and virtual singers have been implemented. Another application that integrates virtual talking faces into emails was also described and implemented.

8.2 Suggestions for Future Works

Several suggestions to improve the proposed system are listed as follows.

(1) Reduction of the number of visemes --- The viseme information of 115 classes of Mandarin syllables is required to create animations containing arbitrary Mandarin syllables. However, the number of required visemes can still be reduced, because mouth shapes of some of them are quite similar. Reduction of the number of required visemes can shorten the processing time of learning dramatically.

(2) Detection of base regions automatically on faces with glasses --- In Chapter 4, a knowledge-based face recognition method is utilized to locate the base region on the face of the first frame. Since the method needs to learn the eye-pair on the face, it does not always work well on faces wearing glasses. If this problem can be solved, then the feature learning process will become fully automatic even on faces wearing glasses.

(3) Integration of emotional expressions --- In this study, emotional expressions are not used because they affects many parts of faces. However, if emotional expressions can be integrated, generated talking faces will become more interesting and natural.

(4) Integration of gestures and body actions --- The integration of gestures and body actions is also a topic that is worth studying. Talking faces with gestures and body actions are more lifelike. The job is easier than the one of integration of emotional expressions, because it involves only base images.

(5) Real-time animation generation --- In the animation generation process, a pre-recorded audio must be inputted and analyzed to get the timing information of syllables in it. If the timing information can be learned in real-time, animations that are synchronized with a speaking person can be generated in real time.

(6) Enhancement of smoothing between visemes --- The movement of the lips produced by the proposed system is still not natural enough. Possible enhancements such as interpolation between successive frames can be studied.

(7) Consideration of syllable collocations --- The viseme of an identical syllable may vary while collocating with different syllables, and this phenomenon should be considered to create more natural animations.

(8) Study of freer video recording process --- In the proposed system, it is necessary for a model to read the entire transcript to create animations starred by the model. If the required features can be gathered from several fragments of videos starred by the model instead of a complete video, the process will be freer and more preferable.

References

[1] J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and W. Swartout, "Toward a New Generation of Virtual Humans for Interactive Experiences," *IEEE Intelligent Systems*, Vol. 17, No 4, 2002, pp. 32-38.

[2] C. Zhang and F. S. Cohen, "3-D Face Structure Extraction and Recognition From Images Using 3-D Morphing and Distance Mapping," *IEEE Transactions on Image Processing*, Vol. 11, No. 11, Nov. 2002.

[3] T. Goto, S. Kshirsagar, and N. M. Thalmann, "Automatic Face Cloning and Animation – Using Real-Time Facial Feature Tracking and Speech Acquisition," *IEEE Signal Processing Magazine*, 2001.

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

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

[6] 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.

[7] S. A. King and R. E. Parent, "Lip Synchronization for Song," *Proceedings of the Computer Animation*, 2002.

[8] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Second Edition, pp52-54.

[9] T. Ezzat and T. Poggio, "Facial Analysis and Synthesis Using Image-Based

Models,“ *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, Oct. 1996.

[10] N. M. Thalmann, P. Kalra, and M. Escher, “Face to Virtual Face,” *Proceedings of the IEEE*, Vol. 86, No. 5, May 1998.

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

[12] T. Ezzat and T. Poggio, “MikeTalk: A Talking Facial Display Based on Morphing Visemes,” *Proceedings of the Computer Animation Conference*, Philadelphia, Pennsylvania, June 1998.