Chapter 1
Introduction

1.1 Motivation

With the advance of high-speed networks and image compression techniques, networked multimedia services prevail nowadays. The voice over Internet protocol (VoIP) technology is evolving, which is useful for making telephone calls using computer networks instead of an analog phone line, and thus IP phones are becoming popular.

Actually, many applications of networked multimedia services have emerged, such as video telephones, video chat-rooms, teleconferencing, videos on demand, distance learning, and web TVs, etc. However, the size of multimedia is still too big to circulate in the network with no delay time. Many people devoted themselves to the researches of improving the situation of network congestion in networked multimedia services. One of the ways to solve the problem is to develop an improved compression technique to reduce the size of multimedia. Another way is to use virtual talking faces instead of real images of people. That is, a virtual talking face that looks like a real human is generated to simulate a user's face and expressions in real time. And only the control points of virtual faces are transmitted to the remote site through networks. So the data size transmitted on the network can be reduced, and the transmission speed can be improved.

The animation of virtual faces and the technique of real-time facial expression tracking have been developed for many years. In general, if we want to extract facial

features to control virtual faces, we can put markers on the face and track them continuously by sensors. After mapping the tracked markers to the virtual face, a virtually animated face with expressions can be generated. But this approach is too expensive and complicated for usual users.

For the reasons of low costs and easy generation, we use web-cameras to capture facial features without using markers in this study. The control points of virtual faces are extracted by facial feature region detection. In order to achieve real-time facial feature tracking, we use existing virtual face models and focus on facial feature detection, such as eyes and mouths, to decrease the calculation work. For some applications, like e-learning and web TVs for children, we choose 2D cartoon faces as face models in this study. Then we transform the feature points detected before into the control points of existing virtual face models.

In this study, we wish to propose an approach to detecting facial features from sequential facial images and establishing a real-time virtual face animation system. In this system, users use only web-cameras and microphones to record their images and voices, respectively. The equipments of this system are simple and cheap for usual users. In addition, we hope that we can combine a real-time virtual face animation system with networks to establish a multi-role network system. Based on this study, it is expected that the generation of virtual announcers, virtual teachers, virtual avatars, and so on, will become easier and more popular in our life.

1.2 Survey of Related Studies

Roughly speaking, many issues about the real-time talking cartoon face are studied in this research. The first issue is how to create a lovely cartoon face model. In

[3] and [4], some methods were proposed to create skeleton-based cartoon faces. A skeleton-based cartoon face is comic-like and created by stroke rendering. A method proposed to animate skeleton-based faces is Litwinowicz et al. [5]. Ruttkay et al. [2] designed a system that lets users create a cartoon face by composing the existing facial components provided by the system. Each component has its own time behavior in the animation. Then the cartoon face is animated by each time behavior in the components. In Chen and Tsai [1], an automatic talking cartoon face generation system was proposed. The system can create a cartoon face automatically from an input image. The cartoon face is animated by an audio-visual mapping between input speeches and lips of cartoon face. Another way to animate the cartoon face in the system is to control the lips of the cartoon face from sequential facial images.

The second issue is how to track the facial features in real-time. A method was proposed to animate virtual faces by adding some markers on faces [6]. By using some image processing techniques, it is possible to track facial features from sequential facial images without markers. Goto et al. [7] designed a complete system that can track many facial features in real-time, like eye, eyebrow, mouth, and jaw. An effective eye-pair detection method was proposed in [12]. A real-time facial tracking for eyes and eyebrows was proposed in [10]. Fabrice et al. [9] proposed a robust facial feature tracking method. The method begins with the detection of a nostril-pair. Then the eyebrows and mouth are detected according the position of the nostril-pair. In [8] and [11], a real-time mouth tracking method was proposed. And Lucey [13] proposed an effective mouth segmentation method.

1.3 Overview of Proposed Method

An overview of the proposed approach is described in this section. First, some definitions of terms used in this study are described in Section 1.3.1. And some assumptions made for this study are listed in Section 1.3.2. Finally a brief description of the proposed method is outlined in Section 1.3.3.

1.3.1 Definitions of Terms

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

(1) **Neutral Face:** MPEG-4 specifies some conditions for a head in its neutral state [12] as follows:

1. Gaze is in the direction of the Z-axis.
2. All face muscles are relaxed.
3. Eyelids are tangent to the iris.
4. The pupil is one third of the iris diameter.
5. The lips are in contact.
6. The line of the lips is horizontal and at the same height of lip corners.
7. The mouth is closed and the upper teeth touch the lower ones.
8. The tongue is flat and horizontal with the tip of the tongue touching the boundary between the upper and lower teeth.

In this thesis, we call a face with a normal expression a *neutral face*.

(2) **Neutral Facial Image:** A neutral facial image is an image with a frontal and straight neutral face in it. It usually is the first frame of a facial image sequence.

(3) **Sequential Facial Image:** Sequential facial images are video sequences of a user's speaking face. In this study, such images are extracted from existing videos or captured from a camera in real-time. The first frame of these facial

images is chosen as a neutral facial image.

(4) **Facial Features:** For detecting facial expressions, we pay attention to several facial features. They are eyes, mouth and the range of head turning.

(5) FAPUs: Facial animation parameter units (FAPUs) are the fractions of distances between some facial features, like eye separation, eye-nose separation, and so on.

(6) **Face Model:** A face model is a 2D cartoon face model with 74 feature points and some FAPUs, including hair, eyebrows, eyes, noses, mouths, and so on.

(7) **Image Feature Points:** Such points are extracted from facial feature regions tracked in sequential facial images.

(8) **Face Model Control Points:** These points are some of the 74 feature points of the face model. They are used to control many features of the face model, like eye opening, lip movement, and head turning.

1.3.2 Assumptions

In real situation, it is not easy to track in real-time faces in a disordered and complicated environment. For real-time tracking, the effects of the environment are varied and uncontrollable. We must make a few assumptions and limits in this study to reduce the complexity of the work and the rate of fault. They are described as follows.

(1) Sufficient light is needed in the environment.

(2) The face is located at the center of each captured image to reduce the rate of miss.

(3) The skew angle of the face does not exceed ±10°.

(4) The head in sequential facial images does not move quickly and zoom.

(5) A user's eyes must be obvious in sequential facial images and not mixed with the hair.

1.3.3 Brief Descriptions of Proposed Method

In the study, three systems are designed: a talking-face cartoon video generation system, a real-time talking cartoon face animation system at a local site, and a real-time talking cartoon face animation system for use on networks.

The first system includes four major parts: a video analyzer, an environment regulator, a facial feature tracker, and a video generator. The video analyzer captures the video stream and the audio stream in an input video. The facial feature tracker tracks the facial feature from the video frame sequences. The face model transformer transforms the image feature points into the face model control points. The talking-face cartoon video generator generates finally the talking-face cartoon videos. A flowchart of the system is shown in Figure 1.1.

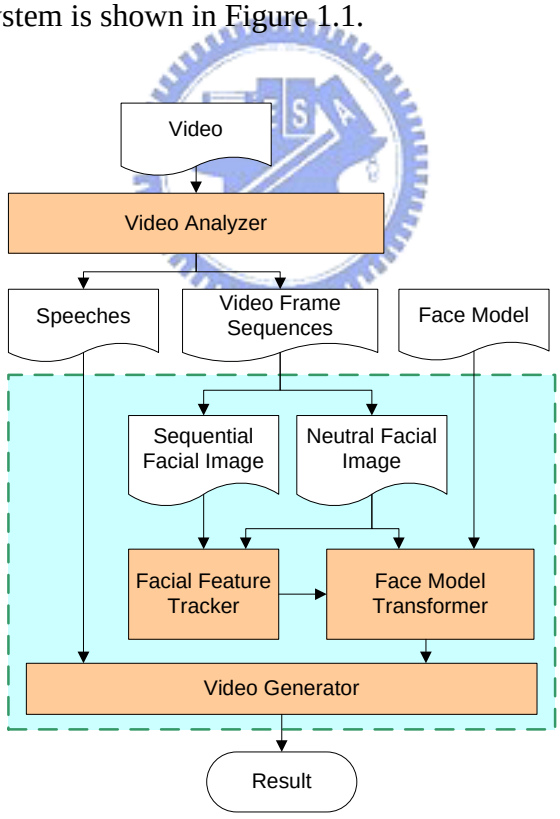

Figure 1.1 A flowchart of talking-face cartoon video generation system.

The second system includes four major parts: an environment regulator, a facial feature tracker, a sound recorder, and an animation generator. The environment

regulator calculates some parameters and lets users regulate them to reduce the miss of the real-time process. The facial feature tracker tracks image feature points in the sequential facial images. The sound recorder records a user's speeches with a microphone. The animation generator generates the talking cartoon face. A flowchart of the system is shown in Figure 1.2.

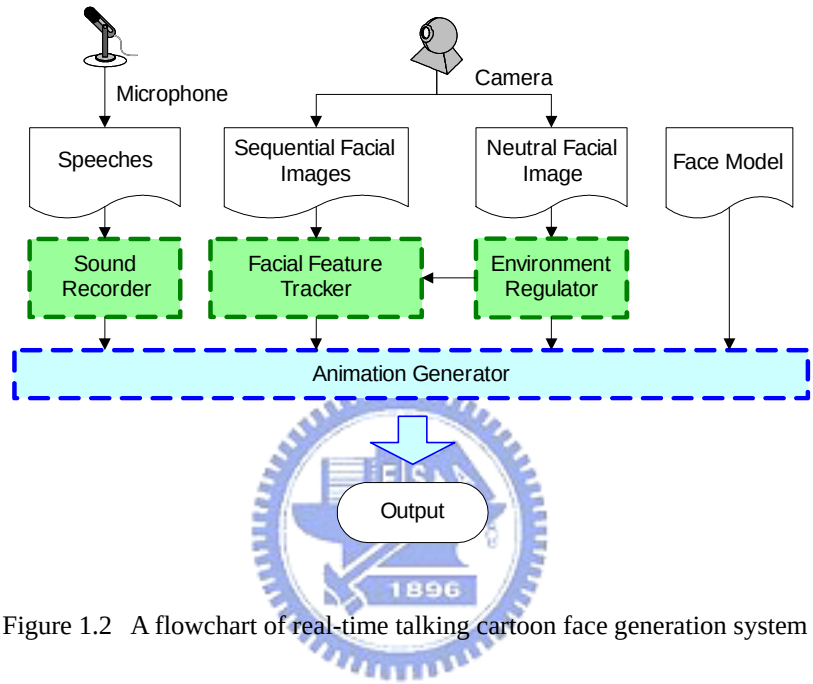

Figure 1.2 A flowchart of real-time talking cartoon face generation system

The third system has two components: a server subsystem and a client subsystem. The server subsystem includes four major parts, an environment regulator, a facial feature tracker, a sound recorder, and a data transmitter. The environment regulator lets users regulate parameters to help tracking. The facial feature tracker tracks the image feature points from the sequential facial images. The sound recorder records the speeches. The data transmitter transmits the image feature points and the speeches to the remote client site through networks.

The client subsystem includes two major parts: a data acceptor and an animation generator. The data acceptor accesses the image feature points and the speech from networks. The animation generator generates the talking cartoon face. A flowchart of the system is shown in Figure 1.3.

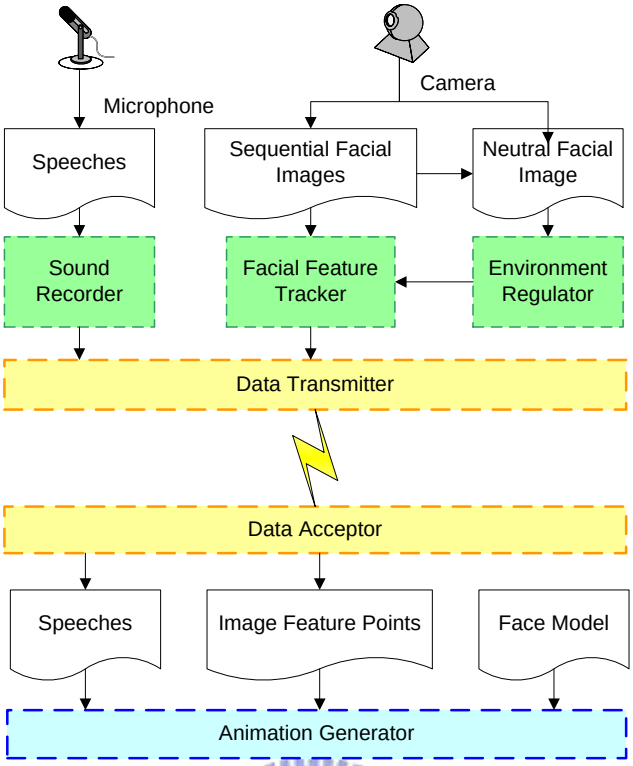

Figure 1.3 A flowchart of real-time talking cartoon face generation system through networks

1.4 Contributions

Some major contributions of the study are listed as follows.

(1) A method of facial feature tracking from sequential facial images in complicated backgrounds is proposed.

(2) A method for transforming image feature points into face model control points is proposed.

(3) A method for detection of head turnings from sequential facial images is proposed.

(4) A system for talking-face cartoon video generation is designed.

(5) A system for generation of real-time talking cartoon faces is designed.

(6) A system of generation of real-time talking cartoon faces for use on networks is designed.

1.5 Thesis Organization

The remainder of the thesis is organized as follows. In Chapter 2, the proposed method of transformation of facial feature points and a method of creation of virtual cartoon faces are described. In Chapter 3, the proposed method of facial feature tracking from sequential facial images is described. Up to Chapter 3, talking cartoon faces are generated from sequential facial images.

The proposed system for generation of talking-face cartoon videos from videos or face feature data is described in Chapter 4. The proposed system for generation of real-time talking cartoon faces is described in Chapter 5. And in Chapter 6, the proposed system for generation of real-time talking cartoon faces for use on networks and some applications are described. Finally, conclusions and some suggestions for future works are included in Chapter 7.