Chapter 4
Automatic Generation of Talking-Face Cartoon Videos

4.1 Introduction

Up to now, we have known how to track image feature points from sequential facial images and how to transform image feature points into face model control points. Therefore, a system for generation of talking cartoon faces from sequential facial images can be designed. In this study, two kinds of systems for generation of talking cartoon faces are designed: one is off-line and the other is real-time.

In the proposed off-line talking cartoon face generation system, the sequential facial images and the speeches are captured from videos. And applying the methods for facial feature tracking and image feature point transformation described previously, talking-face cartoon videos are generated. In the real-time talking cartoon face generation system, the sequential facial images are captured from a camera and the speeches are recorded from a microphone in real time. Then the talking cartoon faces are generated in real time by applying the methods for facial feature tracking and image feature point transformation.

In this chapter, a system for off-line generation of talking-face cartoon videos is proposed using the methods described in Chapter 2 and Chapter 3. Some overview of the file formats used in input videos is described in Section 4.2. A proposed system for off-line generation of talking-face cartoon videos using AVI files as inputs is

described in Section 4.3. Another proposed system for off-line generation of talking-face cartoon videos using the face feature data tracked from sequential facial images and the speeches recorded before as inputs is described in Section 4.4.

4.2 Overview of AVI Video File Format

The file format of input videos used in this study is a Microsoft AVI (Audio Video Interleaved) video file. The AVI file format is a resource interchange file format (RIFF) used to play, capture, and edit audio or video sequences. AVI files contain multiple streams of different types of data, such as videos, audios, and texts, etc. By reading and editing the streams of videos and audios, it is possible to generate talking-face videos from AVI files. An overview of the RIFF format is described in Section 4.2.1. And a detailed specification of the AVI format is described in 4.2.2.

4.2.1 Overview of RIFF Format

The RIFF specified by Microsoft is a file format for multimedia files. RIFF files use FOURCC codes to identify file elements. A FOURCC (four-character code) is a 32-bit quantity representing a sequence of one to four ASCII alphanumeric characters.

The basic block of an RIFF file is called a *chunk*. A chunk is a logical unit of multimedia data and each chunk contains the following fields:

(1) A four-character code specifying the chunk identifier;

(2) A double word value specifying the size of the data member in the chunk;

(3) A data field.

A chunk contained in another chunk is called a *subchunk*. Only two kinds of chunks with a chunk identifier of “RIFF” or “LIST” are allowed to contain subchunks. The

“RIFF” chunk is unique and must be the first chunk in an RIFF file. ALL other chunks in the file must be subchunks of the “RIFF” chunk.

“RIFF” and “LIST” chunks include an additional field in the first four bytes of the data field. This additional field provides the form type of the filed in “RIFF” chunks and provides the list type of the field in “LIST” chunks. The form type is a four-character code identifying the format of the data stored in the file. The list type is a four-character code identifying the contents of the list. An illustration of an RIFF format is shown in Figure 4.1.

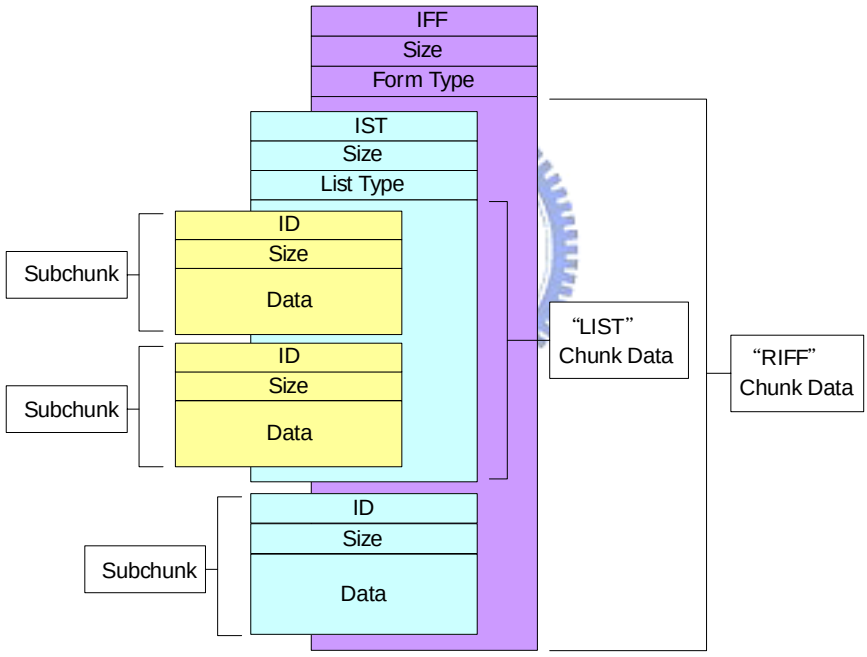

Figure 4.1 An illustration of an RIFF format.

4.2.2 Overview of AVI Format

The AVI file format is based on the RIFF document format. AVI files are identified by the FOURCC “AVI” in the “RIFF” chunk. All AVI files include two “LIST” chunks. The first “LIST” chunk is identified by the FOURCC “hdrl” and defines the format of the data streams. The second “LIST” chunk is identified by the

FOURCC “movi” and contains the data for the AVI sequence. Sometimes an AVI file also includes an index chunk, which provides the location of the data chunks in the file. The index chunk is identified by the FOURCC “idxl”

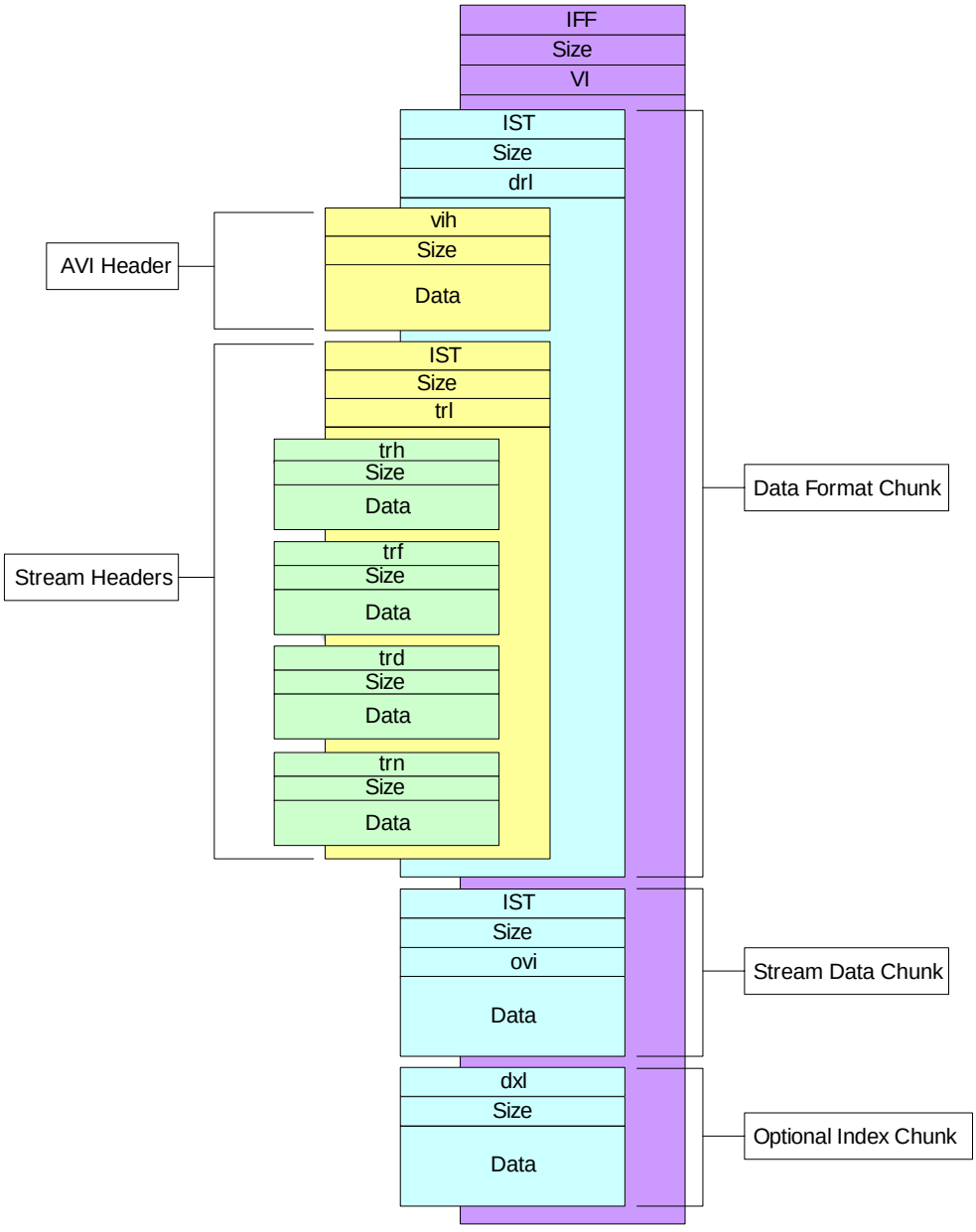

Figure 4.2 An illustration of an AVI format.

The “hdrl” list contains many subchunks. The first subchunk is an “avih” chunk,

which denotes the main AVI header that contains the global information for the entire AVI file. One or more “strl” lists, which contain the information about each stream in the file, follow the first chunk. Each “strl” list must contain a stream header chunk identified by “strh” and a stream format chunk identified by “strf”. Moreover, some “strl” lists might contain a stream-header data chunk identified by “strd” and a stream name chunk identified by “strn.”

The “movi” list that follows the “hdrl” list contains the data in the streams, such as video frames and audio samples. Sometimes an “idxl” chunk can follow the “movi” list. The “idxl” chunk contains a list of the data chunks and their locations in a file. An illustration of an AVI format is shown in Figure 4.2.

4.3 Generation of Cartoon Videos from AVI Files

In this section, the basic idea and the configuration of the proposed method for generation of talking-face cartoon videos from AVI video files is illustrated in Section 4.3.1. And a generation process is described in Section 4.3.2.

4.3.1 Basic Idea and Configuration

If a user wants to generate a talking cartoon face from an existing video, a way to represent talking cartoon faces is to create a corresponding talking-face cartoon videos. In this study, we use the AVI files mentioned in Section 4.2 as the input and output videos. By using a VFW (Video for Windows) SDK provided by Microsoft, the data streams in AVI files can be extracted and edited.

The proposed system for generation of talking-face cartoon videos includes four major parts: a video analyzer, a facial feature tracker, a face model transformer, and a video generator. The video analyzer is used to extract the video and audio streams from an existing video. And the information in the video and audio steams is analyzed to create the output video. The facial feature tracker is used to track the facial features from video frames. The face model transformer is used to transform the image feature points into the face model control points. The video generator is used to generate the finally talking-face cartoon videos. A configuration of the talking-face cartoon video generation system is shown in Figure 4.3.

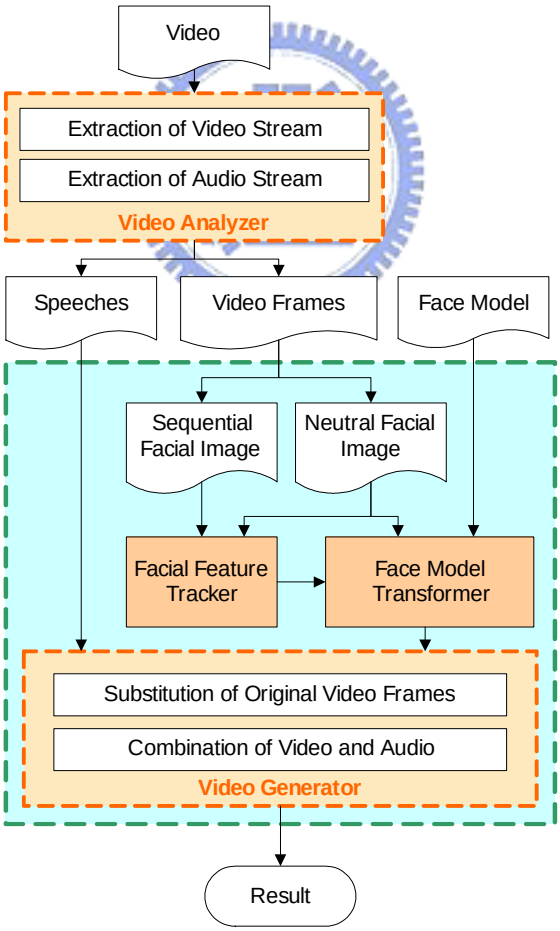

Figure 4.3 A Configuration of the talking-face cartoon video generation system from AVI files

4.3.2 Generation Process

First the video analyzer extracts the video stream and the audio stream from an AVI file. Some information in the video stream is analyzed, such as the length of the stream and the frame rate. The video frames with BMP formats are extracted from the video stream using the VFW SDK even if the video frames are compressed. The first frame in the video frames is taken as the neutral facial image. When the facial features and the FAPUs in the neutral facial image are detected completely, the facial feature tracker begins to track the image feature points from each frame in the video frames. Then the face model transformer starts to transform the image feature points into the face model control points with the face model chosen by a user.

Finally, the video generator renders the sequential cartoon face images instead of the original video frames according to the face model control points. And the new video frames and the speeches extracted from the original video are combined to generate the finally talking-face cartoon video.

4.4 Generation of Cartoon Videos from Face Feature Data in Processed Image Frames

In this section, the basic idea and the configuration of the proposed method of generation of talking-face cartoon videos from face feature data is illustrated in Section 4.4.1. And a generation process is described in Section 4.4.2.

4.4.1 Basic Idea and Configuration

Sometimes a user wants to use an existing video to generate several talking-face cartoon videos with different face models. There is a problem that too much time is wasted by doing the same facial feature tracking in each generation of talking-face cartoon videos. Therefore, another system for generation of talking-face cartoon videos from face feature data is proposed. The face feature data mean the image feature points and the FAPUs extracted from the video frames before. The data are saved in a file.

The proposed system for generation of talking-face cartoon videos from face feature data includes two major parts: a face model transformer and a video generator. The face model transformer is used to read the face feature data in a file and transform them into the face model control points. The video generator is used to generate the talking-face cartoon videos. A configuration of the talking-face cartoon video generation system is shown in Figure 4.4.

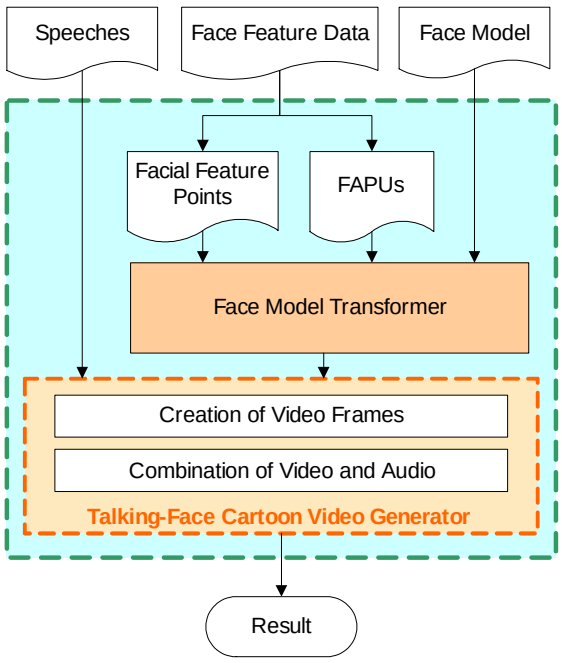

Figure 4.4 A configuration of talking-face cartoon video generation system from face feature data

4.4.2 Generation Process

First, the face model transformer reads the FAPUs and the face model in the files. Then the sequential image feature points read from the file are transformed into the face model control points. Then the video generator generates finally the talking-face cartoon videos step by step as follows.

1. Create a video stream and an audio stream in a new AVI file using the VFW SDK.
2. Render the sequential cartoon face images and assign them into the new video stream.
3. Read the speeches from a WAV file and assign them into the new audio stream.
4. Combine the video stream and the audio stream to create the output AVI file.

4.5 Experimental Results

Some experimental results of the system for generation of talking-face cartoon videos are shown in this section. An input video sequence used here is shown in Figure 4.5. The user was saying two Mandarin words “夢想”. An experimental result of the talking-face cartoon videos generated from the video sequence is shown in Figure 4.6. And the face feature data extracted from the video sequences are saved in a file. An experimental result of the talking-face cartoon videos generated from the face feature data with another face model is shown in Figure 4.7.

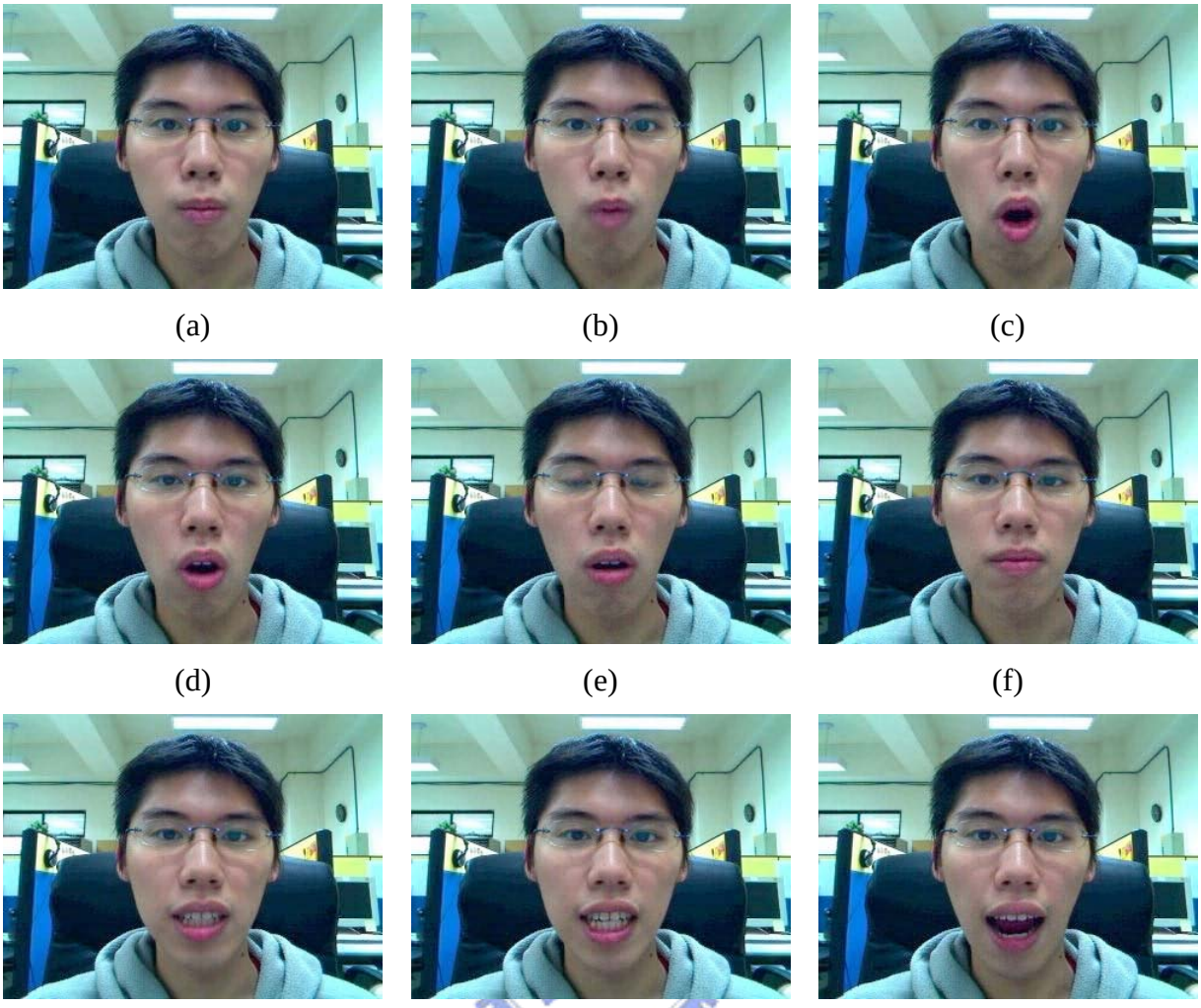

Figure 4.5 An input video sequence used for the experiment.

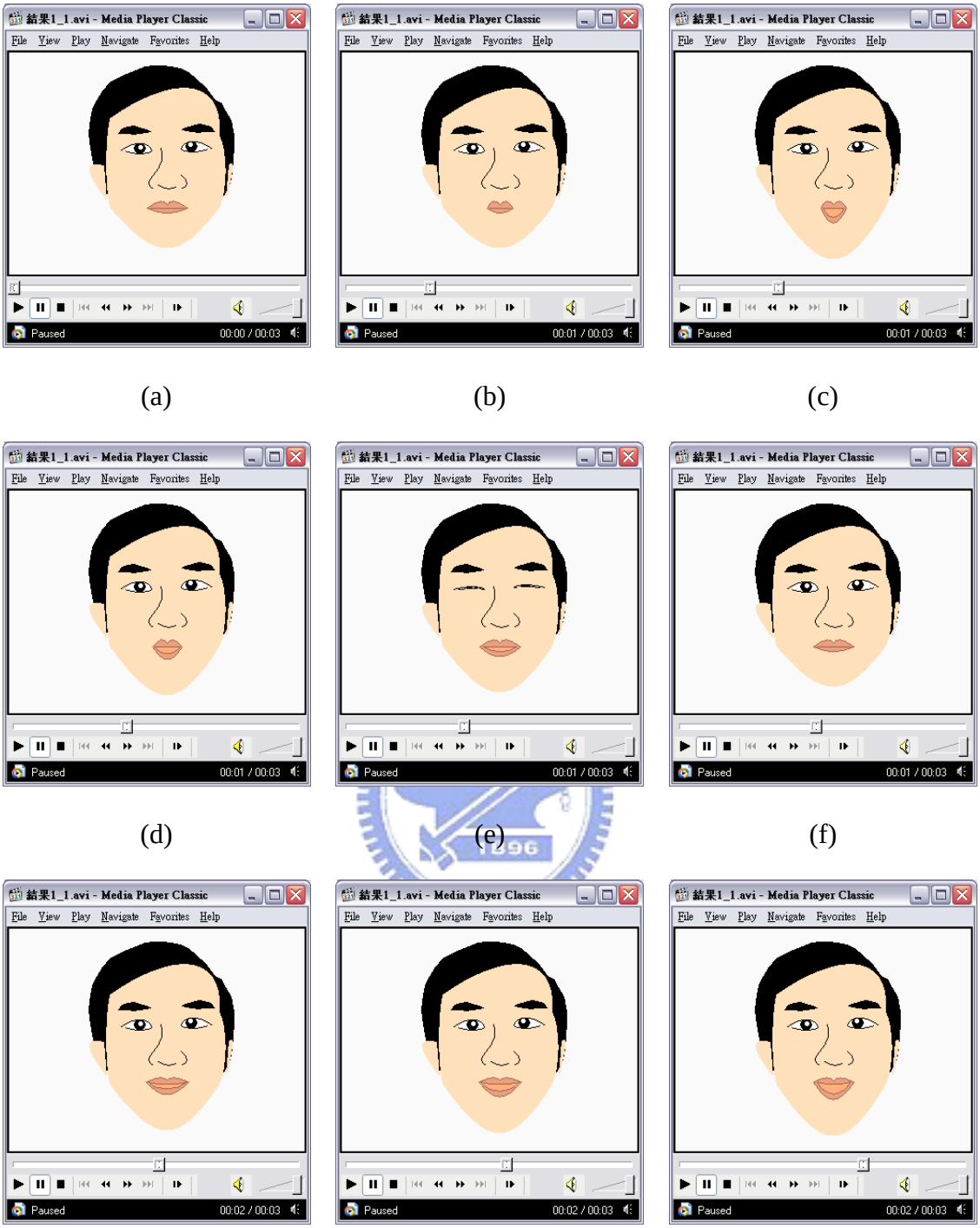

Figure 4.6 A resulting sequence of the talking-face cartoon video for “夢想”, which are generate from the image sequence in Figure 4.5.

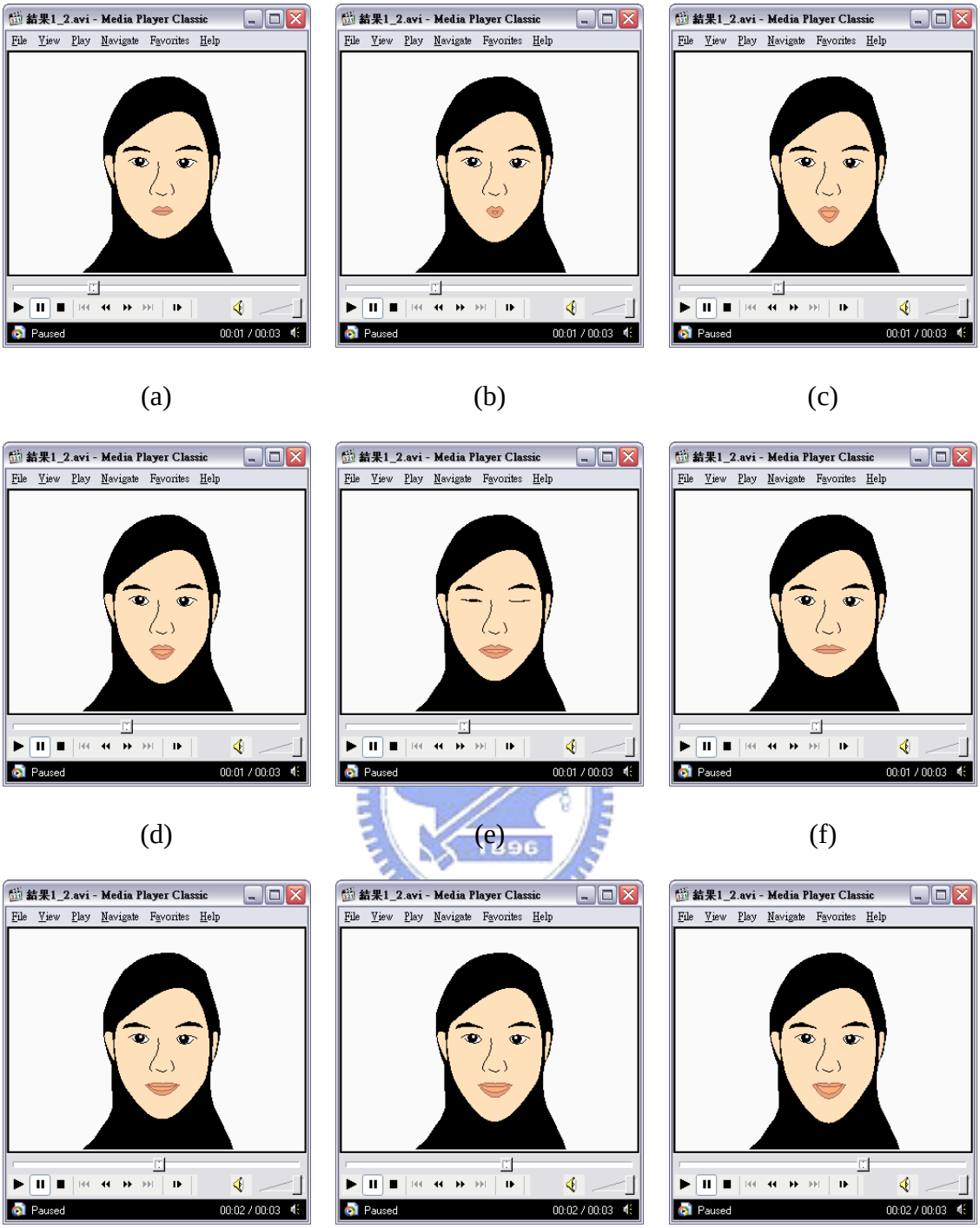

Figure 4.7 Another resulting sequence of the talking-face cartoon video for “夢想”, which are generated by the face feature data extracted from the image sequence in Figure 4.5.