AUTOMATIC REAL-TIME GENERATION OF TALKING CARTOON FACES FROM IMAGE SEQUENCES IN COMPLICATED BACKGROUNDS AND APPLICATIONS

Yen-Long Chen[1] and Wen-Hsiang Tsai[1, 2]
[1] *Institute of Computer Science and Engineering*
National Chiao Tung University, Hsinchu, Taiwan, R. O. C.
[2] *Department of Computer Science and Information Engineering*
Asia University, Wufeng, Taiwan, R. O. C.
E-mail: {gis93562, whtsai}@cis.nctu.edu.tw

ABSTRACT

A system for automatic real-time generation of talking cartoon faces is proposed, which includes five processes, namely, preprocessing, facial feature tracking, image feature point transformation, speech recording, and animation generation. The processes are designed to allow the system to create cartoon faces fast enough for real-time display of the result. A method for creation of oblique 2D cartoon faces is also implemented. Two interesting applications on networks are also implemented, namely, multi-role avatar broadcasting through network, and web TV by the ActiveX technique. Experimental results show the feasibility of the proposed methods.

1. INTRODUCTION

Virtual face animation has been developed for many years [1-5]. Some methods were proposed to animate a virtual face from input facial image sequences with attached markers on faces [6]. If one wants to track facial features without any markers, some image processing techniques can be used [12, 13]. Methods for real-time facial feature tracking have also been proposed [7-10]. In this study, we focus on virtual talking face creation to produce animation which can be transmitted on the Internet in real time.

For low-cost and efficient animation, we use web cameras to capture facial features without using markers. 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 2D virtual face models and focus on facial feature detection, such as eyes and mouths, to decrease the calculation work. We transform the detected image feature points into the control points of existing virtual face models. We have applied the proposed automatic real-time technique to two applications: multi-role avatar broadcasting through networks, and web TV by the ActiveX technique. A major contribution of this study is that we complete automatic generation of talking cartoon faces for real-time applications on computer networks.

In the remainder of this paper, the proposed methods of transformation of facial feature points and creation of virtual cartoon faces are described in Section 2. In Section 3, a method for facial feature tracking from sequential facial images is described. A method for generation of talking-face cartoon videos is described in Section 4. The proposed system for generation of real-time talking cartoon faces is described in Section 5. And in Section 6, the two previously-mentioned applications are described. Finally, conclusions are included in Section 7.

2. CARTOON FACE MODELING BY IMAGE FEATURE POINT TRANSFORMATION

The proposed cartoon face generation system includes two parts: *a facial feature tracker and a face model transformer*. The former tracks facial feature points from sequential images. Before doing this, a face model must be constructed first. The feature points in the image and the control points in the face model are independent. Thus, we may track a man's facial features and use a woman's face model to generate a cartoon face of the woman. A transformation between image feature points and face model control points is needed here. The face model used in [1] is adopted in this study and shown in Fig. 1.

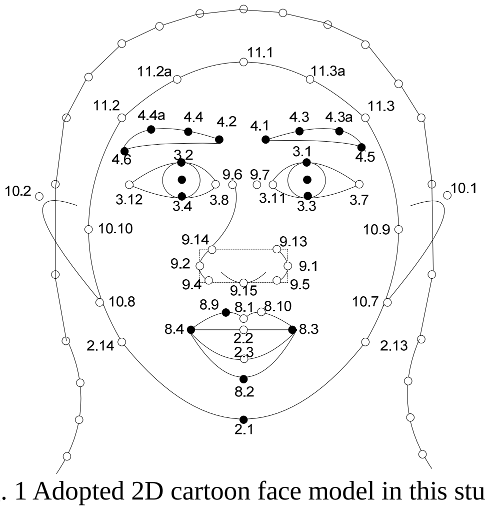


Fig. 1 Adopted 2D cartoon face model in this study.

2.1. Transformation of Image Feature Points into Face Model Control Points

Based on the 2D face model in Fig. 1, we transform image feature points into face model control points by two steps: the first to transform eye feature points and the second the mouth feature points. The idea is to compute the parameters for the transformation according to the ratios of the face model feature values to the image feature values.

The eye features are taken to be the distance h_1 between eye points 3.2 and 3.4 and the distance h_2 between eye points 3.1 and 3.3 shown in Fig. 1. Let the height of an open eye be denoted as H_{eye} which is computed from a neutral facial image. And let H_{eye}' denote the corresponding height in the face model. Then the height h_1' of the open left eye in the face model is computed according to the geometric ratio principle by $h_1' = h_1 \times (H_{\text{eye}}'/H_{\text{eye}})$. The transformation of the right eye is conducted in a similar way using h_2 and h_2'. For mouth point transformation, a *basic mouth point* is defined to be the center of the closed mouth in the face model. Based on the position of the basic mouth point, the mouth points are transformed according to the geometric ratio principle, too. The details are too many to be included here [14].

2.2. Creation of Cartoon Face

Two types of cartoon faces, *frontal* and *oblique* ones, are created by the corner-cutting subdivision and cubic Bezier curve approximation algorithms. They are used to compose a *head-turning* face. A frontal cartoon face is drawn by the 72 feature points of the face model. The static parts of a face, including the hair, face contour, eyebrows, nose, and ears, are created first. Then the dynamic parts of the face, including the eyes and mouth, are created. Two experimental results of creation of frontal cartoon faces is shown in Fig. 2.

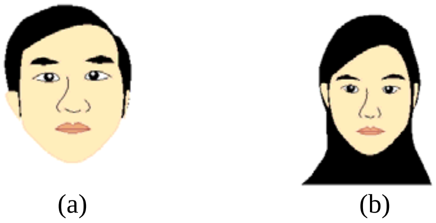


Fig. 2 Experimental results of creation of frontal faces. (a) A male face. (b) A female face.

The basic idea of creation of an oblique cartoon face is to shift some face model control points to change the contour of the frontal cartoon face, resulting in the shifting of some facial features (the eyes, mouth, eyebrows, nose, and ears) as well as the shifting of some contour parts of the face (the cheeks, jaw, and forehead). The face model control points are shifted according to a value t specifying the information about head turning. The sign of t denotes the direction of head turning and the absolute value of t denotes the range of head turning. By using different t values, the facial features may be skewed to represent various degrees of face obliqueness. Two experimental results are shown in Fig. 3.

3. FACIAL FEATURE TRACKING FROM IMAGE SEQUENCES IN COMPLICATED BACKGROUNDS

A new method based on Chen and Tsai [1] is proposed for real-time facial feature tracking in this study. The method can handle more effectively images with shadows, shaky heads, insufficient lighting, complicated grounds, and yielding smooth eye blinking and lip movement results.

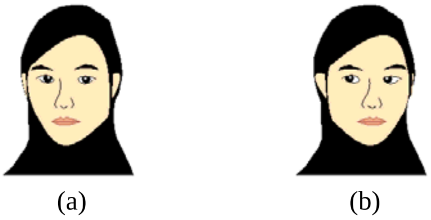


Fig. 3 Two experimental results of creation of oblique cartoon faces. (a) An oblique cartoon face with t = 8. (b) Another oblique cartoon face with t = –8.

3.1. Segmentation of facial feature regions

Chen and Tsai [1] used hierarchical bi-level thresholding to segment facial feature regions, hairs, and backgrounds in sequential facial images. But this is for creating a face model mainly, not for facial feature tracking. In order to let the system more stable and efficient, some methods for segmentation of facial feature regions aiming at facial feature tracking are proposed. In the method for segmentation of eye regions, a central rectangle is used to collect the color information of a user’s face. And the threshold value used to get the binary image is computed in the rectangle. An experimental result is shown in Fig. 4.

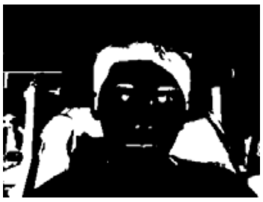


Fig. 4 An experimental result of segmentation of eye-pair regions in intensity channel.

To reduce mouth tracking errors, a method using chromatic color features for segmenting mouth regions is proposed. According to the eye-pair region position, a rectangle that covers the mouth region is speculated. And an octagon included in the rectangle is computed to adapt to the contour of the mouth. Some threshold values are computed by analyzing the normalized red and green intensity values of the pixels in the octagon in a normalized chromatic space. The normalized chromatic space is defined by the following equation based on an image in the RGB space: $r = R/(R+G+B)$, $g = G/(R+G+B)$. The mouth region is segmented out by thresholding the sequential facial images by use of such values in the normalized chromatic space. An experimental result of segmentation of mouth regions is shown in Fig. 5.

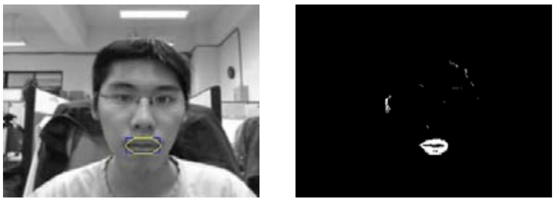


Fig. 5 An experimental result of mouth segmentation of regions. (a) Octagon fitting a rectangle enclosing the mouth. (b) Result.

3.2. Proposed method of eye tracking

In the process of real-time facial feature tracking, first we apply region growing in a tracking window formed by the position of pupils to extract the new eye region. Then the positions of pupils are computed by the Chen and Tsai's eye-pair detection method [1]. And correction of region extraction errors is conducted according to the difference between the positions of pupils in the previous frame and the current frame. An experimental result is shown in Fig. 6.

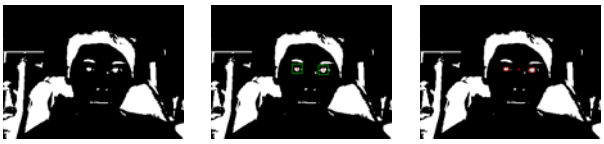


Fig. 6 A result of eye-pair region extraction. (a) Binary image B_{eye}. (b) The $n \times n$ square in B_{eye} with n = 9. (c) The final result of extraction of eye regions.

3.3. Proposed method of mouth tracking

In the extraction of mouth region, we apply region growing to the tracking window formed by the mouth region in the previous frame to extract the new mouth region in the current frame. And based on the position of basic mouth point, four mouth points are extracted. To avoid mouth tracking misses, some corrections are conducted. First, we correct a type of *absolute error* showing that the mouth region is too small to represent a normal mouth. Second, we correct a type of *relative error* indicating that the mouth regions have an unreasonable change in size between the two consecutive frames. An experimental result is shown in Fig. 7.

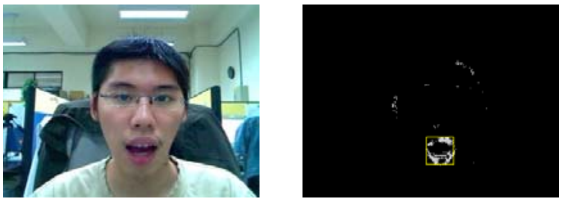


Fig. 7 A result of mouth region extraction. (a) Original image. (b) Final result of mouth region tracking.

3.4. Detection of head turning

The distance between the cheeks and the center of the mouth changes with the range of head turnings, thus some values are defined for detection of head turnings. Let D_{LfCheek} and D_{RtCheek} denote respectively the distances between the center of the mouth and the left and right cheek. We obtain the degree of head turning R_{turn} by computing the ratio of D_{LfCheek} to D_{RtCheek} using the pixels in the Sobel edge image.

4. REAL TIME CARTOON FACE ANIMATION

In a system for real-time talking cartoon face generation, four functions are necessary. First, it is basic to process image sequences in real time. Second, the work for recording and playing speeches in real time is similarly indispensable. Third, the performance of real-time facial feature tracking is easily affected by uncontrolled environment. So a friendly interface between the user and the system for real-time generation of talking cartoon faces is required. Finally, integration of videos and audios is needed.

4.1. Organization of proposed system

The proposed system of real-time talking cartoon face generation includes four major parts: *a facial parameter regulator*, *a facial feature tracker*, *a sound recorder*, and *an animation generator*. The facial parameter regulator automatically extracts facial parameters from a neutral facial image and provides a friendly interface for easy regulation of these parameters to reduce real-time process misses. The parameters needed are the threshold values for segmentation of eye-pair regions and mouth regions and a division line. The facial feature tracker is used to track image feature points from the sequential images, which are captured in real time. The sound recorder is used to record a user's speeches in real time with a microphone. The animation generator is used to generate talking faces by rendering the cartoon face image and synchronizing the video and the audio in real time.

4.2. Preprocessing of Real-time Tracking

Before starting the real-time facial feature tracking from sequential facial images, some parameters must be confirmed first. In the method for off-line facial feature tracking, the neutral facial image is the first frame in the image sequence. The features extracted from the neutral facial image are used for tracking in the image sequences. But in the method for real-time facial feature tracking, the neutral facial image of the image sequence is unknown until tracking is started. Therefore, a preprocessing of real-time facial feature tracking is necessary. Then relevant features can be extracted from the neutral facial image in the image sequence to reduce errors after tracking is started.

In the preprocessing step, a user can capture several neutral facial images and analyze these images to achieve the optimal results. Then these resulting parameters are averaged to obtain the finally parameters for real-time facial feature tracking.

4.3. Real-time cartoon face generation process

After starting track facial features in real time, the first image captured from a camera is taken as the neutral facial image. If the features are extracted from the neutral facial image correctly, the facial feature tracker keeps capturing the image sequence from a camera and tracking the image feature points. Then, the animation generator transforms the image feature points into the face model control points and renders the cartoon face images in real time.

Sometimes an exception might happen in the process of real-time facial feature tracking, like the wrong size of tracked mouth regions and the unreasonable change in mouth region sizes between the two consecutive frames. The proposed facial feature tracker is designed to ignore the exception by using the facial feature points in the previous frame to recover the facial feature points in the frame that has an exception. If the exception is very grave, the facial feature tracker will stop the process and notify the users that an exception has happened. The entire process for real-time generation of cartoon faces is shown in Fig. 8.

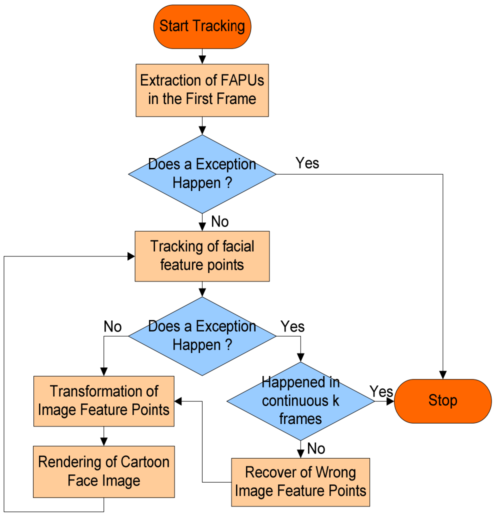

Start Tracking
Extraction of FAPUs in the First Frame
Does a Exception Happen ?
Yes
No
Tracking of facial feature points
No
Does a Exception Happen ?
Yes
Transformation of Image Feature Points
Happened in continuous k frames
Yes
Stop
No
Rendering of Cartoon Face Image
Recover of Wrong Image Feature Points

Fig. 8 Flowchart of proposed real-time cartoon face generation.

In order to capture a user's speeches, a Waveform Audio SDK provided by Microsoft is used in this study. The speeches are recorded with a simple waveform-audio format specification: mono, 8.0 kHz, and 8 bits per sample. Although the quality of digital sounds is not the best, it cannot be discriminated by the human's ear. An audio buffer is used when recording the speeches. If the audio buffer is full with the speeches captured by a microphone, the output audio device plays the speeches stored in the buffer immediately and releases the data in the buffer. By repeating these works, a system for real-time talking speeches can be established.

If a user wants to stop the system for real-time talking speeches, the system will keep recording speeches until one audio buffer is full, and then the recording is stopped. And the output audio device will play the speeches in the full audio buffer and stop.

In most studies of talking virtual faces, the syllables of speeches are analyzed for synchronizing moving lips and speeches. However, the methods for analysis of syllables are different for different languages. In this study, we do not analyze the syllables of speeches captured from a microphone. The lip movements of talking cartoon faces are synthesized according to the input real sequential facial images. It is possible then to generate talking cartoon faces in real time and synchronizing with any speeches.

On the other hand, automatic synchronization of cartoon videos and speeches is a difficult problem because no information about speeches can be used. In order to solve this problem, a friendly interface is proposed to help users to synchronize the cartoon videos and speeches. The basic idea is to fix the timing of speeches and regulate the delay of images to synchronize the cartoon videos and speeches. In this interface, a delay value is used to control the delay of images. The delay value denotes the delay of frames in the system for real-time generation of talking cartoon faces. A user can change the delay value to regulate the delay of images to synchronize the cartoon videos and speeches in real time.

Although a speech recognizer, such as the Microsoft Speech SDK, can be integrated into the system to achieve lip sync automatically, yet it can only deal with three languages (Chinese, English, and Japanese). Our method is more general to handle all kinds of languages.

5. APPLICATIONS OF REAL-TIME CARTOON FACE ANIMATION TO MULTI-ROLE AVATAR BROADCASTING AND WEB TV

Talking cartoon faces can be used on networks as web tutors to help students study on networks. they also can be used as broadcasters to announce messages through networks. A real-time cartoon face animation system for such uses on networks is proposed.

5.1. Organization of Proposed System

The proposed system is composed of two subsystems: a server and a client. By interaction between them, the video and audio data in the server can be transmitted to the client in real time. The server subsystem includes four major parts, a facial parameter regulator, a facial feature tracker, a sound recorder, and a data transmitter. The facial parameter regulator, the facial feature tracker, and the sound recorder are the components of the real-time cartoon face animation system. They are used to generate the image feature points and the speeches. The data transmitter is used to transmit the image feature points and the speeches to the remote client site through networks. The image feature points are transmitted when the facial feature tracking is completed in each frame, and the speeches are transmitted when the audio buffer is full.

The client subsystem includes two major parts: a data acceptor and an animation generator. The data acceptor is used to access the image feature points and the speeches from networks. The audio output device plays the sound when the speeches are accessed. When the image feature points are accessed, the animation generator generates the talking cartoon face in real time with a face model chosen at the client site. A configuration of this system is shown in Fig. 9.

5.2. Video and Audio Transmission

In video transmission, the image feature data in each frame are packed as a *video unit* with an order number. There are two kinds of video units in the transmission. The first one is the *header unit*, which contains features extracted from the neutral facial image and is used to help the client subsystem transform the image feature points into the face model control points. The second is the *frame unit* which contains the values of the image feature points in a frame.

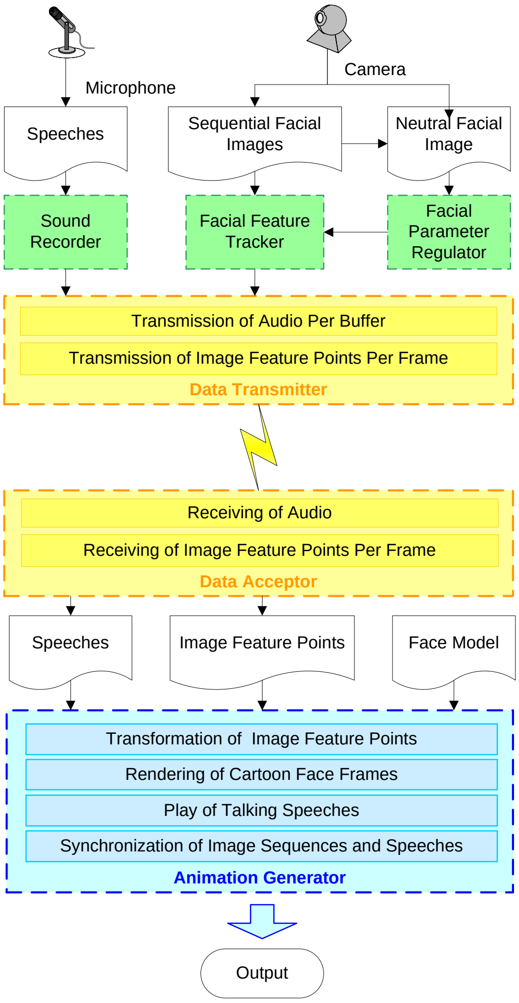

Microphone
Camera
Speeches
Sequential Facial Images
Neutral Facial Image
Sound Recorder
Facial Feature Tracker
Facial Parameter Regulator
Transmission of Audio Per Buffer
Transmission of Image Feature Points Per Frame
Data Transmitter
Receiving of Audio
Receiving of Image Feature Points Per Frame
Data Acceptor
Speeches
Image Feature Points
Face Model
Transformation of Image Feature Points
Rendering of Cartoon Face Frames
Play of Talking Speeches
Synchronization of Image Sequences and Speeches
Animation Generator
Output

Fig. 9 Configuration of real-time cartoon face animation system.

At the beginning, the server transmits the header unit to the client site. Then, the client accesses the parameters in the header unit. If the access process is successful, the client returns a value to notify the server to start transmitting the frame units sequentially.

In the audio transmission, the data in an audio buffer are packed as an *audio unit* with an order number. The server starts to transmit the audio unit to the client when video transmission starts. The client accesses the audio data in the audio unit and plays the audio data.

5.3. Application to Multi-role Avatar Broadcasting through Networks

The proposed multi-role avatar broadcasting system allows a receiver on the network to choose different avatars to animate talking cartoon faces according to the broadcaster's images and speeches. It can be used in the public place of entertainment. The avatars can be different at different receptor sites of the broadcast system. So the broadcast system can be variegated. An illustration is shown in Fig. 10.

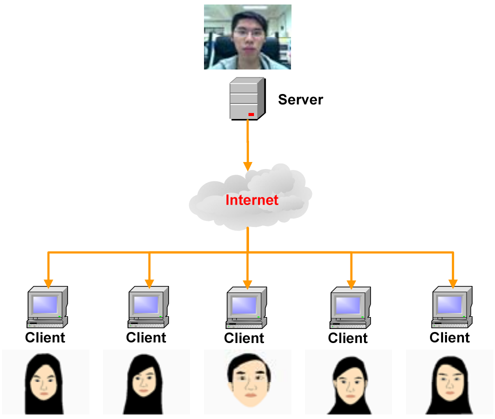


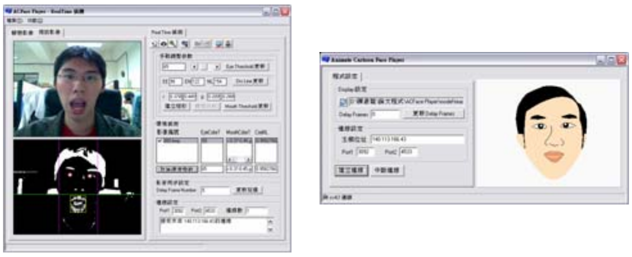
Fig. 10 An illustration of multi-role avatar broadcasting.

In multi-role avatar broadcasting, first a broadcaster regulates the facial parameters at the server site. Then he/she waits for the connections of the receivers. Each receiver chooses a face model as the avatar and connects to the server. The broadcaster then starts to broadcast messages to the receivers through networks. The server tracks the facial features of the broadcaster and records his/her speeches. Then the video and audio data are transmitted to each client site. The talking cartoon face is rendered in real time at the client site. If the images and speeches are asynchronous, the receiver changes the delay to synchronize them. An experimental result is shown in Fig. 11.

5.4. Application to Web TV by ActiveX Technique

Another application of the server and client system is the web TV, which we implemented in this study. A user can use an IE browser to receive the messages from the server. Via the interface of web TVs, avatar tutors, avatar reporters, and avatar singers can be broadcast to the world through networks. An illustration of the web TV is shown in Fig. 12.

The process of creating a web TV is similar to the process of multi-role avatar broadcasting. The difference is that the operations at the client site are in a web browser. An activeX technique is used here to implement the client system. An activeX control is downloaded automatically from the server and executed by a web browser at the client site. A user must enter the URL of the server to start this process. An experimental result is shown in Fig. 13.



(a) (b)

Fig. 11 An experimental result of multi-role avatar broadcasting system. (a) The server system. (b) The client system.

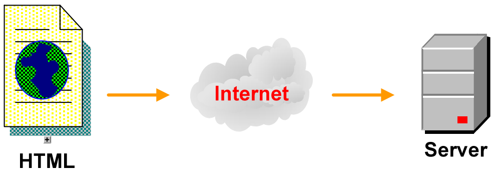


Fig. 12 An illustration of the web TV.

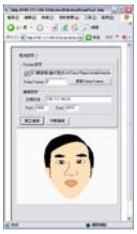


Fig. 13 An experimental result of web TV.

6. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORKS

In this study, a system for automatic real-time generation of talking cartoon faces has been implemented, which is based on analysis of 2D sequential facial images. The system consists of three components: a facial feature tracker, a face model transformer, and an animation generator. A head turning detection method was proposed to create head-turning cartoon faces. And a method was proposed to synchronize the images and speeches in real time. In addition, a server and client system for application uses on networks has been proposed. By the integration of the server and client subsystems, two applications were be implemented, namely, multi-role avatar broadcasting and web TV. Some interesting topics for future research include raising the quality of the generated talking cartoon faces and the accuracy of real-time facial tracking.

REFERENCES

[1] Y. L. Chen and W. H. Tsai, “Automatic Generation of Talking Cartoon Faces from Image Sequences,” *Procs. 2004 Conf. on Computer Vision, Graphics & Image Processing*, Hualien, Taiwan, August 2004.

[2] Z. Ruttkay and H. Noot, “Animated CharToon faces,” *Procs. of 1st Int’l Symposium on Non-photorealistic Animation and Rendering*, Annecy, France, June 05-07, 2000, pp. 91-100.

[3] H. Chen, N. N. Zheng, L. Liang, Y. Li, Y. Q. Xu, and H. Y. Shum, “PicToon: a personalized image-based cartoon system,” *Procs. of 10th ACM Int’l Conf. on Multimedia*, France, 2002, pp. 171-178.

[4] R. Hsu and A. K. Jain, “Generating discriminating cartoon faces using interacting snakes,” *IEEE Trans. on Pattern Anal. & Machine Intell.*, 2003, vol. 25, pp. 1388 -1398.

[5] P. Litwinowicz and L. Williams. “Animating images with drawings,” *SIGGRAPH94 Conf. Proceeding*, Addision Wesley, August 1996, pp. 225-236.

[6] D. Burford and E. Blake, “Real-time facial animation for avatars in collaborative virtual environments,” *Procs. of South African Telecommunications Networks and Applications Conf.* 1999, pp. 178-183.

[7] T. Goto, S. Kshirsagar, and N. Magnenat-Thalmann, “Real Time Facial Feature Tracking and Speech Acquisition for Cloned Head,” *IEEE Signal Processing Magazine*, May 2001, vol. 18, no. 3, pp. 17-25.

[8] G. C. de Silva, T. Smyth, and M. J. Lyons, “A Novel Face-tracking Mouth Controller and its Application to Interacting with Bioacoustic Models,” *Procs. of 2004 Conf. on New Interfaces for Musical Expression*, June 03-05, 2004, pp. 169-172.

[9] F. Bourel, C. C. Chibelushi, and Adrian A. Low, “Robust facial feature tracking,” *Proc.s of 11th British Machine Vision Conf.*, Bristol, UK, 2000, pp. 232-241.

[10] A. Kapoor and R. W. Picard, “Real-Time, Fully Automatic Upper Facial Feature Tracking,” *Procs. of 5th Int’l Conf. on Automatic Face & Gesture Recog.*, May 20-21, 2002, pp. 10-15.

[11] N. Oliver, A. P. Pentland, “Lafter: Lips and Face Real Time Tracker,” *Procs. of IEEE Conf. on Computer Vision & Pattern Recognition*, Puerto Rico, 1997, pp. 123-129.

[12] S. C. Y. Chan and P. H. Lewis, “A Pre-filter Enabling Fast Frontal Face Detection,” *Procs. of Third Int’l Conf. on Visual Information and Inform. Systems*, June 02-04, 1999, pp. 777-784.

[13] S. Lucey, S. Sridharan, and V. Chandran, “Adaptive mouth segmentation using chromatic features,” *Pattern Recognition Letters*, September 2002, pp. 1293-1302.

[14] Yen-Long Chen, “Automatic Real-time Generation of Talking Cartoon Faces from Image Sequences in Complicated Backgrounds And Applications,” *Master’s Thesis*, Institute of Computer Science & Engineering, Nat’l Chiao Tung University, Hsinchu, Taiwan, June 2006.