A Software Facial Expression Synthesizer with The Chinese Text-to-Speech Function

Woei-Luen Perng, Yungkang Wu, Ming Ouhyoung

Communication and Multimedia Laboratory,
Dept. Of Computer Science and Information Engineering,
National Taiwan University, Taipei 106, Taiwan

ABSTRACT

The proposed system is named Image Talk: a real-time synthetic talking head using one single image with Chinese text-to-speech capability. Image Talk uses one single image to automatically create video-like talking sequences in real time. The image can be acquired from photographs, video clips, or hand drawn characters. This interactive system accepts Chinese text input and talks back in Mandarin Chinese, generating facial expression in real time.

Image Talk analyzes Chinese text by converting it to a standard Pinyin system used in Taiwan and fetches the associated facial expressions from an expression pool dynamically. The expressions are synchronized with the synthetic speech and played back in the video-like talking sequence in real time.

Image Talk also incorporates eye blinking, small-scale head rotation and translation perturbations, to make the results more natural. The generic Talk Mask is also easy to switch to any other facial or non-facial images, such as dogs, for special effects. The result is quite entertaining, and can easily be used as a new human-machine interface, as well as for lip synchronization for computer animated characters.

Keywords: Facial Animation, Image Warping, Real-time Lip Synchronization, Texture Mapping.

1 INTRODUCTION

Deceased man talks, is that possible? From the current techniques in computer vision and graphics, it is possible. You have seen it in movies, and you will see it in Image Talk (see **Figure 1**.)

1.1 Overview

The proposed system is named Image Talk: a real-time synthetic talking head using one single image with Chinese text-to-speech capability. Image Talk gives life to a single static image. It applies a generic Talk Mask to a given image. The character in the image can blink eyes, move its head, and talk in Mandarin Chinese. It accepts Chinese text input and transforms it into internal command that fetch the expressions and then synchronizes with associated sound, thus making a static image talk like a real human being.

Image Talk synchronizes the facial animation with audio speech automatically without labor-intensive human interference. The only human interference is the hand-labeled key frames done while developing the software of Image Talk. We use a generic Talk Mask so that when the system is running, there is no need to edit the key frames again. For this implementation, only nine key frames were chosen as reference frames for mimicking the mouthing of Mandarin speaking. Once Image Talk is running, any new image can be applied in real time with a little adjustment of Talk Mask to match the major features on the face.

The Talk Mask is a generic mask that can be applied to different images. Users only have to modify the mask to match with major facial features, such as eyes and lips, to make another image talk. The Talk Mask was not produced based on Dr. Sun Yet-Sen's (®]¤ ¤ }) facial image, but the results in **Figure 1** are still convincing.

Image Talk has an internal Chinese text-to-speech system, which accepts Chinese input and generate synthetic speech with facial animation at the same time. The speech is generated from collections of natural utterance from a real human. The system is built for real-time experiment with the lip synchronization of the facial animation.

Examples such as Mona Lisa (the painting of Leonardo da Vinci, see **Figure 2**), President Lee Teng-Heui(§ õ ½)(**Figure 3**), and Mayor Chen Shuei-Bian(³ ¤ ô)ó show good results. The animated lip motion with the sound extracted from the speech of the original person makes Image Talk quite believable.

Image Talk can also be applied to non-human

Figure 1: The left most one is the original image of the founding father of modern China, Dr. Sun Yet-Sen, and the rest images are generated from Image Talk. Note that the rightmost Dr. Sun has his eyes closed. The head is moving and blinking eyes dynamically while the system is running.

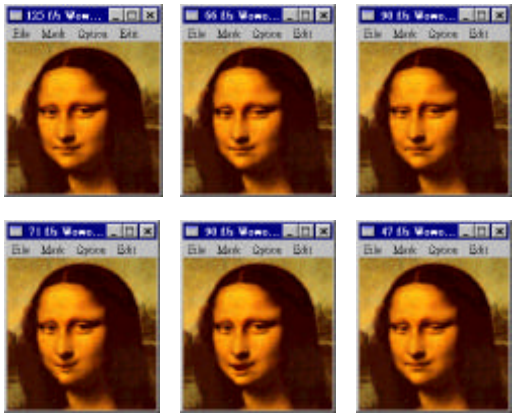

Figure 2: The upper left is the original Mona Lisa. The generic Talk Mask changes her facial expressions in real time.

Figure 3: Examples from Image Talk. The upper row is the original images of the late and the current president. Note that the second row shows when they blink their eyes.

characters, such as dogs. The possibility is only limited by one's own imagination. So we can make Mona Lisa uttering a man's voice, or a dog talk like Mayor Chen Shuei-Bian by using his sound clips. This technique is also very useful for the animation of cartoons, computer animated characters, and other special effects to be used in educational programs, TV commercials, and movie. The model diagram in **Figure 4** shows the process of Image Talk.

1.2 Related Work

Psychologist Ekman and Friesen proposed the Facial Action Coding System (FACS, 1978) that distinguishes all possible visually distinguishable facial movements. In 1980's [Waters87] proposed a more general and flexible muscle model for parameterization that will allow facial control without the requirement for hard-coding the performable actions. Adding photographic texture onto a face proved to have good result such as in [Oka87]. In 1990's, with the advanced hardware provided by companies like Cyberware, more realistic three-dimensional facial models were done by [Williams90] and [Lee95].

Image metamorphosis has proven to be a powerful tool for visual effects. This process, commonly known

ional geometric image warping with color interpolation. Image transitions before the developments of morphing were generally achieved through the use of cross-dissolves, which is linear interpolation to fade from one image to another, but it does not retain the geometric analogy of the two. The problem was solved by applying two-dimensional mesh warping with the process of color interpolation. [Beier92] provides a more expressive way to achieve a similar result. The mapping of mesh now reduces to the mapping of features, thus named feature-based image metamorphosis. This method would requires more computing complexity but with more desirable result.

For telecommunication at very low bit-rate, some model-based coding methods were employed in videoconferencing [Lavagetto94, Aizawa 95]. The 2 1/2 dimensional facial model is dynamically adapted to time-varying facial expression by means of few parameters, estimated from the analysis of the real image sequence. By transmitting the encoded parameters only, the large bandwidth needed for transmitting video sequence in videoconferencing can be reduced dramatically.

In Video Rewrite [Bregler97], Bregler uses audio

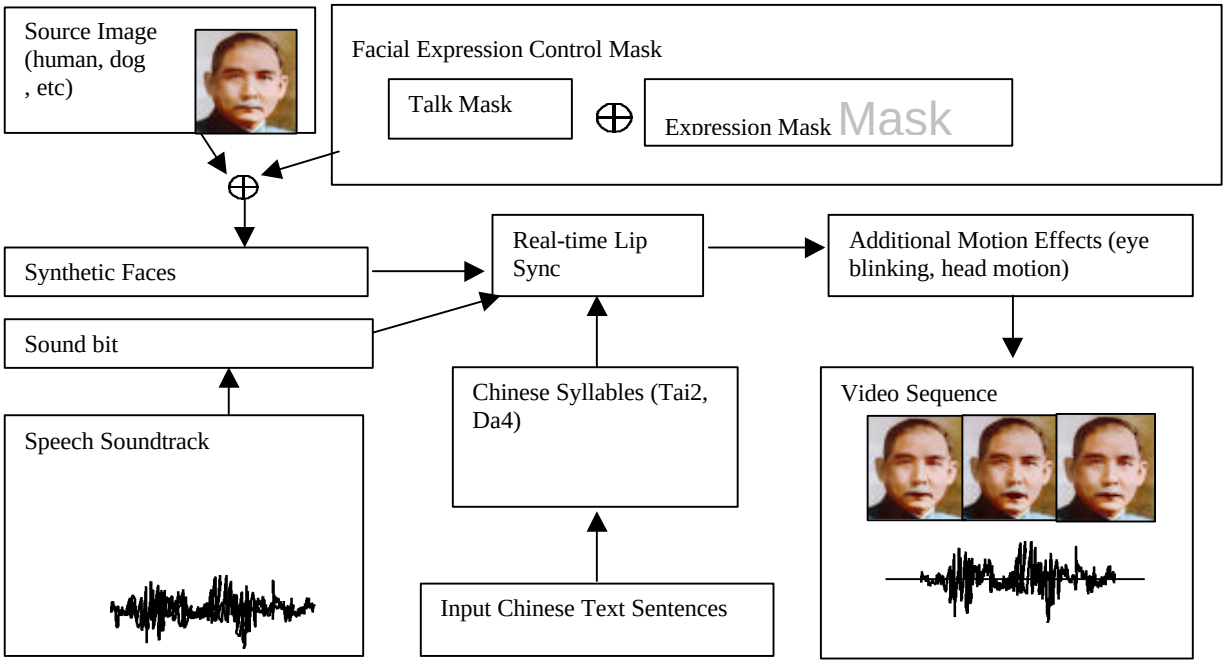

Figure 4: Model Diagram of Image Talk

track to segment the training video sequence into triphones, and then select from this video model to synthesize the new lip motion for to any given audio. The triphone includes key elements of facial features in talking, eyes, mouth and chin. This technique is useful in movie dubbing where the movie sequences can be modified to synchronize the actor's lip motions to the new soundtrack.

Cartoon animators use key frames for all the action of the characters and have to draw the intermediate frames by hand laboriously. The facial expressions were in general done by artists' imagination with the references to one's own facial expression reflected in a mirror or photographs and video sequences taken by a camera in advance. Many other applications could have been discovered by using a talking head as an interface for text-to-speech system. The British Telecom proposed a three-dimensional talking head to be the front end of its text to speech synthesizer, Laureate, and DECFace from DEC has similar idea. MikeTalk from MIT (Videorealistic Talking Faces: A Morphing Approach) uses the visemes (extracted from training video sequence) associated with English speech to generate talking sequence from a string of phonemes generated from the text-to-speech system. There is even entertainment software, Virtual Friends (HapTek, Inc.), released in the first half of 1998.

2 Facial Animation in Image Talk

Image Talk models a human head by applying a two-dimensional mesh model, and uses real-time mesh warping for animations. Since it is simple, it can perform in real-time. We introduce a cheap and simple procedure to model a realistic talking head. The talking head is supposed not to have much head rotation and translations. Inspired by the 2 1/2 dimensional model used in [Lavagetto94], we focussed the research solely on a two-dimensional mesh model developed from a static facial image. The color information of a neutral face provides a basis image for the system.

A neutral face is the frontal facial image without specific facial expressions. By manipulating the image, we can morph the neutral face into various expressions. A set of interesting spot is marked as control vertices. These vertices were placed around the eyes, nose, mouth and chin. The control vertices were then connected into primitives, which in this case were triangles. The convex shapes, such as triangle, would make texture mapping simple.

For our work, less then 200 control vertices were marked on the facial area, and the final two-dimensional mesh model comprises less than 350 triangles (See **Figure 5**). The mesh model is normalized into a generic mask, the Talk Mask, therefore can be applied to other facial image too.

2.1 Synthetic Facial Expression

The texture applied to the two-dimensional meshes is the original texture coordinate with the neutral facial expression image. Affine transformations are used in texture mapping. The first step of facial animation is to define the key frames. The neutral face without any

expressions can be seen as a key frame that contains a neutral facial expression. We manipulate the texture-mapped image to edit specific facial expression based on the images from video clips and the reflection of the author in a mirror. The image warped in real time as the control vertices were adjusted. The key frames are saved as the vector difference of each control vertex from the neutral facial expression, and normalized according to the size of the generic mask.

Synthetic face can be represented as linear combinations of generic Talk Mask and Expression Mask.

FacialExpression = GenericTalkMask + ExpressionMask

The expression vectors are recorded by the mask coordinate system. Differences between the manipulated control vertices and the neutral mask vertices were stored.

2.2 Time Driven Interpolation

Image Talk makes the face talk in real time by interpolating the key frames in real time. Since the expressions are normalized vectors according to a generic mask, for time driven interpolation, the transitional facial expressions for the real time morphing from *Expression A* to *Expression B* used in this implementation is a linear combination of the two.

3 Mouthing Mandarin Chinese

The most important issue in Image Talk is to make the computer-animated character speak Chinese. Image Talk generates talking sequences in real time. The sequences are based upon a few collections of key frames of facial appearances. Image Talk matches the key frames with the Chinese syllables and produce smooth visual speech automatically (see **Figure 6**)..

3.1 Mandarin Chinese Syllables

Chinese is a tonal language. A typical syllable may have up to five different tonal variations, and they represent different words, or different meanings. Many of the Chinese words were pronounced with the same syllable or even with the same tone, and syllables could be described by Pinyin system. The Pinyin system called "Ju4 In1 Fu2 Hau4"(ª `- ¾¹ Å¸)¹ comprises symbols to describe the sound of Chinese syllables. According to this method, Mandarin Chinese pronunciations are classified into 408 monotone syllables [Lee89]. There are 1333 pronunciations if the tonal variations are also considered.

Figure 5: This image is a good example of a neutral face. It is the basis image where the Talk Mask is developed. The origin (0.0,0.0) is set to the upper-left of the image. The image coordinate system sets the lower-right corner as (1.0, 1.0). The base and size of mask is recorded for a specific image.

In order to make the talking head look like speaking Chinese, we have to devise an efficient way to build the visemes, which are the key frames needed for eachsyllable. We approach it by utilizing the characteristic of Chinese syllables, which are the sound elements every student in Taiwan has to learn in the elementary school, the Pinyin method. All the 408 Chinese syllables can be described by 37 Pinyin symbols. These symbols are what we learned about the standard Chinese pronunciation in Taiwan. Pinyin symbols can be categorized into initials and finals.

From the observation of natural speech video sequences, we can find that to pronounce a Chinese word, one has to shape one's lip from the preparation of initials to the end of finals. To produce specific sound correctly, the mouth has to form a specific shape, and by moving the tongue inside the mouth cavity with the vibration of the vocal cord in the throat. The Pinyin system is a convenient and efficient way to describe the mouthing shape of the syllables. The observation implies that the visemes of syllables can be composed of the movement from initial to final.

3.2 Static Key Frames

After observing video clips of persons speaking Chinese, we found that there are basic lip shapes for pronouncing Chinese utterance. Most of them are strongly tied with the shape of lips while pronouncing individual Pinyin symbols. We selected and saved the extreme feature images from the footage of a real person speaking pronouncing Chinese words. Usually one extreme feature image for the initials, and one to two extreme-feature images for the finals. We classified them into some basic facial key frames for each initial and final. Many initials and finals may refer to the same key frames as their shape of lip does not make much difference in talking sequences. The implementation of the static facial expressions of the initials comprises only one key frame, and of the finals comprise some one to four frames. The total facial expressions defined are less then 10 frames, which are all the key facial expressions for the pronunciation of Chinese syllables.

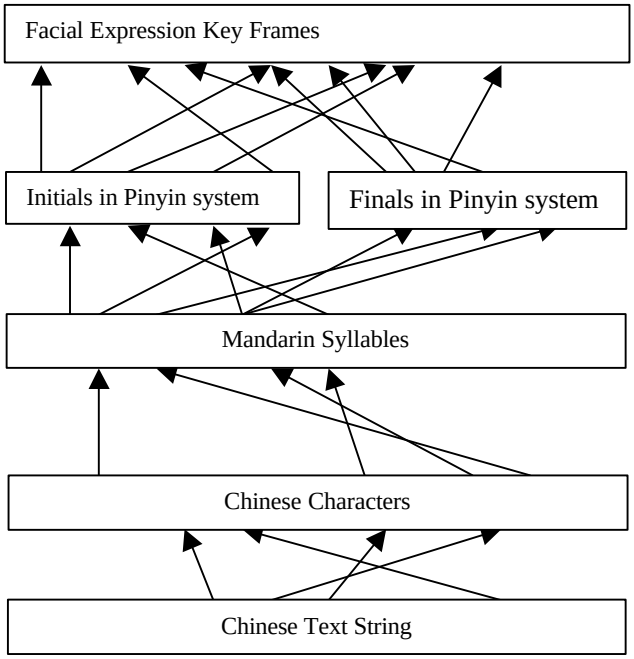

Figure 6: A systematic way to match key frames with Chinese syllables. As user input Chinese text string, it is converted into Chinese characters, then match to specific syllables. From the syllable, we can choose the key frames from the composed initials and finals.

3.3 How To Build Visemes from Key Frames

Image Talk emphasizes on the lip shape for uttering Mandarin Chinese, the teeth and tongue were ignored at the current experiment. To pronounce a Mandarin Chinese syllable, one does not always form the lip shape from initials to finals. For syllables started with plosive sound, such as "bo, po, mo, fo", it is true that one has to prepare the shape for pronouncing the initials. But for other initials, the shapes for them are mostly dissolved into the shape of finals. The rules are applied while selecting key frames to build viseme. The viseme-building process does not finish here. To make the viseme animate with the Chinese speech, we have to devise a smart method to synchronization both of them.

4 SPEECH SYNCHRONIZATION

Image Talk retrieves the syllables after analyzing the input sentence. The syllables were synthesized using the animation techniques described earlier, to produce a smooth animation with corresponding sound. The results would be like a talking head giving a speech.

4.1 Synchronize With a Syllable

During the input process, the Chinese characters transformed into syllables and then mapped to specific Pinyin elements, next the visemes are fetched. The viseme contains one or several basic key frames for pronouncing the syllables. From the observations of video clips of real talking sequences, we can use specific proportion of time duration between the key frames to mimic the animated sequence of speech. There are some rules used in our system:

(1) In order to make the sound of syllable, the mouth have to make a specific shape before pronouncing any sound.

(2) The syllable can be pronounced later, even after some time the mouth has shaped in advance.

(3) The mouth shape can be kept for some time even after the sound of syllable has finished.

These rules imply that a sound has to be defined by key facial expressions. The two obvious parameters needed are the time for preparing and ending a syllable. The starting parameter is the time from the end of last word to the actual sound pronounced of from current word. The ending parameter is the time for pronouncing the syllable. To fulfill the first and second rules described above, the shape of mouth has to be ready before the starting parameter. As for the last principle, the shape of mouth can keep in the same position after the ending parameter, or started to the next desired facial expression. (see **Figure 12**)

If a syllable have the time duration of **ad** (see **Figure 12**), and ***n*** facial expressions, ($\mathbf{F_1}$, $\mathbf{F_2}$, ..$\mathbf{F_n}$), before time **b**, the facial expression has to morph to $\mathbf{F_1}$ already. Furthermore, the morphing sequence has to be finished before time **c**. Note that the ending parameter does not mean the sound of a syllable has ended, but refers to the end of a viseme.

For this implementation, we set the $\mathbf{F_1}$-$\mathbf{F_n}$ to start and finish at exactly time **b and** time **c**. For the frames during time **b-c,** which is the time to morph from $\mathbf{F_1}$-$\mathbf{F_n}$, we let them to the evenly divided for smooth transition. Assume the time for a syllable is normalized to be 1.0, the staring parameter **b** is set to 0.2 and **c** is set to 0.6 for this implementation.

4.2 Synchronize with Continuous Speech

For continuous speech synchronization, the facial expression before time **a** and after time **d** will not go back to the neutral face in most cases. They should be preparing for the next syllable or just try to move back to the neutral face. We have two choices (a)move facial expression back to a semi-neutral face, (b) set the frame before time **a** to be the last frame ($\mathbf{F_n}$) of the previous syllable sequence, and set the frame after time **d** to be the first frame of the next syllable sequence.

For method (a), we can define the facial expression has a contraction force turning the mouth back into neutral expression. After time **c** of the current syllable, the time **c** to **d** is used to contract back to semi-neutral expression. For method (b), the $\mathbf{F_1}$ is morphed from

the last frame of last syllable. Next, the $\mathbf{F_n}$ will be morphed to the first frame of the next syllable. This makes the mouth continuously prepared for the next syllable as the speech is continuously given.

5 SYSTEM IMPLEMENTATAION

Image Talk is implemented on a Pentium PC with Windows 95/NT. The synthesizer has several internal managers; it is an integration of a simple text-to-speech synthesizer and a talking head system. Image Talk speaks three to five words a second. The speed is determined by the collection of the sound bits.

5.1 System Configuration

The image width and height can be scaled to desired width and height. The texture mapping is done by standard OpenGL Library, SGI's implementation. The performance on the Pentium 166 personal computer is around 20 frames per seconds by software when the window size is about 200 pixels by 300 pixels in true color mode. The performance reaches over 80 frames per second while the machine is equipped with a hardware accelerator for OpenGL, such as WinFast L2200 (LeadTek ,Inc.) The size of the window can be adjusted as desired by standard pointing device input, such as a mouse. The system would run faster as the window size decreases and slower as the window size increases.

5.2 Text-to-Speech

To enable Image Talk interactively accept and talk different Chinese sentences, a text-to-speech system is developed. Text-to-speech system has to transform the text into syllables and play the sound.

5.2.1 Synthetic Speech. To make the talking head announces Chinese speech; we have to synthesize the Chinese syllables with the facial animation. To make things easier, in this implementation we do not synthesize them artificially; we collect them and recombine syllables into new speeches.

For the experiments of natural speech, we also have test data collected from public-domain which are the normal speech addressed by Mayor Chen Shuei-Bian (http://www.taipei.gov.tw), and Dr. Sun Yet-Sen (you can acquire his audio clips at http://peacock.tnjc.edu.tw/ROC.HTML). Those sounds of syllables are of variable length but still can be reconstructed into new speech. The synchronization method described earlier made promising results for these speeches. Different sentence can be pronounced by recombination of the text input. Therefore, we can make a person speaks what he did not say from his own original sound clip. With the soundtrack and facial image of Dr. Sun Yet-Sen, we can recreate the scene that he is alive and addressing to the audience. Therefore, as the audio and visual data of Mayor Chen Shuei-Bian combine in Image Talk, you can see that he greets you in person, or speaks what he did not address before.

5.2.2 Process Chinese Text. The system accepts a series of Chinese characters in big-5 coded mode. The inputted Chinese sentence could comprise any Chinese character provided with the standard Windows 95/NT operating system. Image Talk transforms the Chinese characters to the corresponding syllables, and then dissects the syllables into initials the finals. The Pinyin symbols associated with the Chinese characters are built from the Chinese input method file. The visemes would be formed dynamically by selected key frames after the system analyzes the initials and finals. The key frames associated with the initials and finals are retrieved from an expression pool, and the selected key frames are then morphed in real-time according to the mechanism described earlier to synchronize with the sound.

Before playing the sequence, the sound element has to be retrieved first. As we play the sound, the smooth motion of synthetic face are generated at the same time by the synchronization method described earlier. The sound is played one syllable after the other, with the visual facial expression morphing in real time; the static image is now talking.

5.3 Generalized Mask and Motion Effects

The synthetic images are generated by the control of Talk Mask. Besides controlling the mouth movement, we also use Talk Mask to make special effects such as eyes blinking and head motion.

To make the mesh model more practical, we normalized the mask into a generic one, which can be applied to other different images just by a few adjustments to match the mask to features of a new face. The normalization is done by recording the Expression Masks in the coordination system of normalized Talk Mask. The Expression Masks were recorded as deviations from the standard mask (The positions of control vertices rested in the neutral face), and were resized according to the size of the Talk Mask if the mask was to be matched to a new face. A user can change the image and interactively adjust the Talk Mask to match features of a new face in real time. Image Talk uses the Talk Mask coordinate system to form the Facial Expression Control Mask (which is the combination of Talk Mask and Expression Mask) to synthesize the new facial expressions by texture-mapping the original image to the warped mesh.

The matching of the Talk Mask to the feature process could be done within seconds if one wanted to get an

approximate match by quickly adjust the bounding box of the Talk Mask. Image Talk provides global and local adjustment tools. Users can resize the bounding box to get a quick match of the features, or use the drag-with-force function (**Figure 7**) to locally adjust the minor area inside the bounding box. To map the eyes and lips more accurately, a user may take about one minute to adjust the mask interactively, and verify with the results in real time.

To make the image more natural, head motion, eye blinking and even head position perturbation are introduced.

5.3.1 Head Motion. The head motion is implemented by adjusting the mask with linear transform algorithm described in Figure 7, drag-with-force, to be moved to the left, right, up and down to make the head simulate a small amount of rotation. The control mask does the motion by moving the control vertices in proportion to a specific direction simulating the results of head motion. It looks fine during the swing rotation, but not as good in the nodding rotation.

5.3.2 Eye Blinking. Semi-periodic eye blinking are added to make the face looks like live human. Image Talk does not really process the eyes right now. Blinking eyelids are done by pulling the upper facial area of the eyes. The triangles inside the eyes are arranged to be upside down and are texture mapped as the eyelids pulled down. This looks fine in normal blinking speed that is about 100-200 millisecond per blink.

5.3.3 Perturbation. To make the image more natural, some perturbation were added to disturb the movement of the mask producing some shape and light variations on the image to simulate a video sequence. The perturbation applied is around 1 to 3 pixels around the control vertices, and makes the shape dynamically change.

6 RESULTS AND CONTRIBUTIONS

Image Talk produces a video-like talking sequence from a single image. When applied to other images, we can even make a deceased man talk. The eye-blinking, head motion, and lip-sync combined with Mandarin Chinese speech gave a good result.

These effects such as small-scale head rotation, eye-blinking play very important rules in making animated character natural. If Image Talk moved only the lip while the character is speaking Chinese, the result looks awkward. Therefore, the system applied random head rotation and eye blinking, perturbations even if there is no Chinese input for synthesizing visual speech. According to our experience, most users are interested in seeing the head moving and eyes blinking even when the character is not talking. It is the key features that Image Talk gives life to a static image.

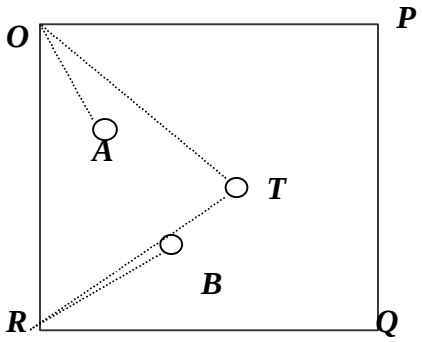

Figure 7: Assume that the left is the bounding box of the Talk Mask. When drag-with-force function is on: if the user is dragging vertex ***T***, and ***A*** is to the upper left of ***T,*** Image Talk keeps the ratio of ***OA*** and ***OT*** to be of the same while ***T*** is being manipulated. For ***B***, which is to the lower left of ***T***, the ratio is kept for ***RT*** and ***RB***. Vertices to the upper right, and lower right of ***T*** are moved according to the reference to ***P*** and ***Q*** respectively.

6.1 Results and Possibilities

Image Talk does head motion, eye-blinking and perturbation even when there is no input to simulate a live talking head. Users can interactively change the subject by incorporating different images. The hair and ears move with the Talk Mask because we morph the mesh outside Talk Mask, too.

6.1.1 TV broadcasting. After applying the image and soundtrack of the Mayor Chen, the static image jumps into live talk. The result looks like a TV broadcast, or videoconferencing, with the head in focus and addressing to the user.

6.1.2 Lip-reading. Lip-reading helps human to recognize speeches in a noisy environment. While incorporating the image and historical speech by Dr. Sun Yet-Sen, we wake the diseased man up. However, the soundtrack was recorded on May 30, 1924, which was quite noisy and unclear. The lip motions of Image Talk help people to understand what he uttered.

6.1.3 Animal Talk. Special effects make animals, such as dogs, pigs, and horses, talk in films. Can Image Talk make non-human characters talk? We test it on an image of a dog. The face of the dog resembles the human face, so the Talk Mask needs not to make major adjustment. Using soundtrack from Mayor Chen, the dog greets your visit to Taipei City.

6.2 Future Work

Image Talk uses the combination of one single image

and a 2D-mesh model to accomplish sophisticated results. However, there are many issues that can be improved for Image Talk in the future: the viewpoint of eyesight, head motion, and the speech-driven lip synchronization.

6.2.1 Head Motion. The two-dimensional mesh model is just a primitive experiment, which gives us the confidence to redo the experience by incorporating a three-dimensional model. The head rotation looks fine in the "swing" rotation, but not quite well in the "nodding" rotation. A 3D-model can be applied to solve the problem.

6.2.2 Synthetic Organs. Eyelid: Image Talk does not really process the eye part right now. The blinking eyelids were done by pulling the upper facial area of the eyes. The eyeballs can be singled out to make synthetic ones, in order to change the viewing direction of the eyes. In a videoconference system, the eyesight movement captured on the other end can be reproduced to the user. The tongue and teeth should be synthesized by a generic one base on the given image.

6.2.3 Improved Text-to-Speech System. The Chinese Text-to-speech system can be easily replaced by other commercial products. The fidelity of the system would definitely be improved if this interface were integrated with a commercial text-to-speech system.

6.2.4 Speech-driven Lip synchronization. Incorporating speech-reading techniques, Image Talk will become a speech-driven system. We can segment the speech and recognize the vowels automatically, the lip synchronization can be driven by the speeches. The speech may not be limited to Mandarin Chinese, but may be of English, French, German, Spanish, or any sound as along as we have built associated lip shape models.

6.3 Contributions

Image Talk uses a single image to automatically create video-like talking sequences in real time. The image can be acquired from a photograph, video clip, or hand drawn characters. This interactive system accepts Chinese text input and talks back in Mandarin Chinese, generating facial expression in real-time.

Image Talk also incorporates eye blinking, small-scale head rotation and translation perturbations, to make the resulting sequence more natural. The generic Talk Mask makes it easy to switch to any other face or non-facial images, such as animals or household commercial products for special effects. The result is quite entertaining, and can easily be used as a real-time human-machine interface, very low bit rate videoconferencing, or off-line production of movie dubbing, lip synchronization in computer-animated characters, and special effects.

BIBLIOGRAPHY

[Aizawa95] K. Aizawa, Thomas S. Huang. "Model-based image coding: adcanced video coding techniques for very low bit-rate applications." Proceedings of the IEEE. 83(2), pp.259-271, February 1995.

[Beier92] Thaddeus Beier, Shawn Neely. Feature-based image metamorphosis. Computer Graphics, 26(2), pp.35-42, 1992. ISSN 0097-8930

[Bregler97] Christoph Bregler, Michele Covell, Malcolm Slaney. "Video Rewrite: Driving Visual Speech with Audio." Computer Graphics Proceeding (SIGGRAPH 97), pp.353-360, 1997.

[Lavagetto94] Fabio Lavagetto, Sergio Curinga. "Object-oriented scene modeling for interpersonal video communication at very low bit-rate." Signal Processing: Image
Communication 6, pp.373-395, 1994.

[Lee89] Lin-Shan Lee, Chiu-Yu Tseng, Ming Ouh-young, "The Synthesis Rules in a Chinese Text-to-Speech System", IEEE Trans. on Acoustics, Speech and Signal Processing. Pp.1309-1320. Vol.37,No.9, 1989.

[Lee95] Yuencheng Lee, Demetri Terzopoulos, Keith Waters. "Realistic Modeling for Facial Animation." Computer Graphics Proceedings (SIGGRAPH 95), pp.55-62, 1995.

[Litwinowicz94] P. Litwinowicz, L. Williams. "Animating images with drawing." SIGGRAPH 94, Orlando, Fl, pp.409-412, 1994. ISBN 0-89791-667-0

[Oka87] Massaki Oka, Kyoya Tsutsui, Akio Ohba, Yoshitaka Kurauchi, Takashi Tago. "Real-time manipulation f texture-mapped surfaces." ACM Computer Graphics (SIGGRAPH 87), 21(4), pp.181-188, 1987.

[Waters87] Keith Waters. "A Muscle Model for Animating three-dimensional Facial Expression." ACM Computer Graphics (SIGGRAPH 87), 21(4), pp.17-24, 1987.

[Williams90] Lance Williams. "Performance-Driven Facial Animation." ACM Computer Graphics (SIGGRAPH 90), pp.235-242, 1990.