

Image Talk: A Real Time Synthetic Talking Head Using One Single Image with Chinese Text-To-Speech Capability

Woei-Luen Perng, Yungkang Wu, and Ming Ouhyoung

Communication and Multimedia Laboratory,

Dept. Of Computer Science and Information Engineering,

National Taiwan University, Taipei 106, Taiwan

Abstract

Image Talk uses a single image to automatically create 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 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. The expressions are synchronized with the sound and played back in a talking sequence.

Image Talk also incorporates eye blinking, small-scale head rotation and translation perturbations, to make the resulting sequence more natural. It is also easy to switch to any other face images. The result is quite entertaining, and can easily be used as a new human machine interface, as well as for lip sync in computer animated characters.

1 INTRODUCTION

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

Image Talk gives life to a single static human facial image. It first applies a generic Talk Mask to a given image. Then the character in the image can blink eyes, move its head, and even talk in Mandarin Chinese. It transforms Chinese text into internal commands that fetch the expressions and sync with associated sound, thus making a static image talk like a real human being.

The Talk Mask is a generic mask that can be applied to different images. The users only have to modify the

mask to fit major facial features, such as eyes and lips, to make another image talk. For example, the Talk Mask was not produced based on Dr. Sun Yet-Sen's facial image, but the result in Figure 1 is still convincing.

1.1 Previous Work

Computer based facial expression modeling and animation is not new, but has attracted more and more attention for the last two and half decades. In the early 1970's, Parke created a three-dimensional facial animation for the first time[11]; Gillenson developed an interactive system to assemble and edit line drawn facial images; later, Parke developed a parameterized three-dimensional facial model[12]. In the late 1980's, Waters, Magenat-Thalmann developed different muscle based model. In the 1990's, we had seen movies and TV commercials utilizing computer animated characters as key storytelling components, such as the recent film "Toy Story"(PIXAR) and "Dragon Heart"(ILM).

In addition to entertainment, facial animation can be applied in many fields. A natural interface for text-to-speech synthesizer or an aid

for lip-reading makes deaf people more easily to recognize speeches. In very low bit rate teleconferencing, utilizing model based coding system aroused attention in recent years.

1.1.1 3D Model with Texture Map. Different approaches had been tried to achieve lifelike human head. By applying key framing used by traditional animations, the three-dimensional model can generate different facial expressions. Nevertheless, it requires the complete specification of a model at each extreme, or at least the storage of the differences between facial positions. The shading effect without facial skin on it (which is approached by texture mapping later) makes it looks like a

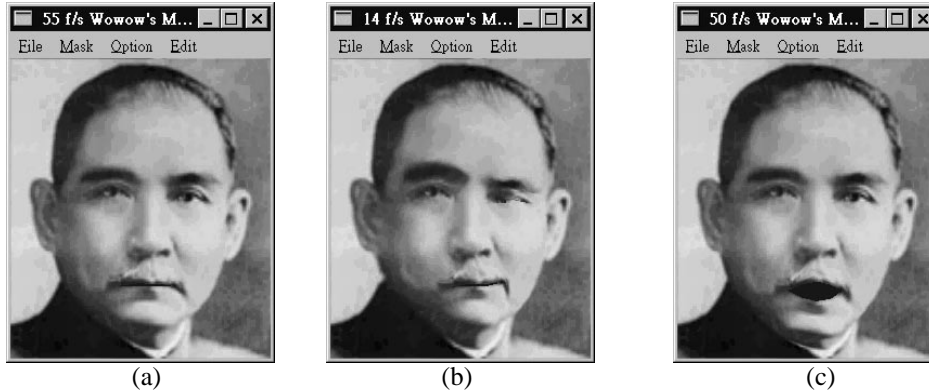


Figure 1: (a) is the original image of the founding father of modern China, Dr. Sun Yet-Sen, and (b)(c) are generated from Image Talk.

doll trying to smile like a human. The later parameterized 3D facial model avoids the problem of rigidity by grouping vertices together to perform specified tasks. However, the generality is lost as soon as the process is applied to a new facial topology.

Psychologist Ekman and Friesen proposed the Facial Action Coding System (FACS, 1978) that distinguishes all possible visually distinguishable facial movements. In 1980s Waters[9] proposed a more general and flexible muscle model for parameterization that will allow facial control without the requirement for hard-coding the performable actions. Generating real good-looking wire framed or shaded facial expression usually did not make a synthesized face realistic. Adding photographic texture onto a face proved to have good result such as in Oka[8].

In 1990's with the advanced hardware provided by companies like Cyberware, more realistic three-dimensional facial model is done by Williams[10] and Lee[6].

1.1.2 2D Image Metamorphosis. Image metamorphosis has proven to be a powerful tool for visual effects. This process, commonly known as morphing, is realized by coupling two-dimensional geometric image warping with color interpolation. Image transitions before the development of morphing were generally achieved through the use of cross-dissolves, which is linear interpolation to fade from one image to another. This problem was solved by applying two-dimensional mesh warping with the process of color interpolation.

Mesh warping was pioneered at Industrial Light & Magic(ILM) by Douglas Smythe for use in the movie "Willow" in 1988 to maintain the two-dimensional geometric correctness. Two key images must have the same amount of meshes that mapped with each other. The process is done by transforming the mesh from one to the other as the interpolation of color is on the way, thus

providing the geometric correctness between the two desired transformed images. The result was good, but it requires cumbersome mapping work between two meshes.

Beier[2] 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.

1.1.3 Videoconferencing. For telecommunication at very low bit-rate, some model-based coding methods were employed in videoconferencing [4][1]. 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.

These model-based encoding has the assumptions that videoconferencing would be two person face to face in front of the terminal just like what we usually see on the television news. The use would have his head and shoulder part shown on the screen and the background are most likely to be static, since the camera would be usually setup at the top the video screen.

Since most of the actions involved in the videoconferencing will be two people talking to each other by using facial expression and rotation of neck and shoulders, the transmitted data between the two would be dramatically reduced by using model-based encoding between two ends. That is to say, the movement of eye and mouth and facial expression can be encoded by a small number of parameters.

1.1.4 Movie Dubbing. In Video Rewrite [3], Bregler uses audio 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.

1.2 Current Applications

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. 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 Overview of 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. To obtain a realistic three-dimensional model, we have to scan a real human face, both the range and image data. The machine used by Williams[10] and Lee[6] would cost quite a fortune and may not be available everywhere. Taking pictures from a specific person at different angles, and then apply to a generic three-dimensional model seems a good idea, too. However, registering the textures from the two-dimensional data to the three-dimensional model would be manually possible but tedious.

Due to economic reason, 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 by Lavagetto[4], we focussed the research solely on a two-dimensional mesh model developed from a static facial image. The facial expressions are then done on the mesh model.

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. The forehead would better not to be covered by hairs, and the chin would better be without beards. If the picture taken was not from a neutral face, the shadows caused by squeezing the muscles while smiling or frowning would sometimes make the synthesized expression unnatural.

A set of interesting spots are 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.

How many control vertices should we mark on the facial area? For real time processing, the less number of the vertices is, the better the performance is. 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. The control vertices were aligned to the contour of specific facial features such as nose, eyes, eyebrows, inner and outer boundary of mouth, the forehead below the hair, and the chin. See Figure 2.

The mesh model is normalized into a generic mask, the Talk Mask, therefore can be applied to other facial image too.

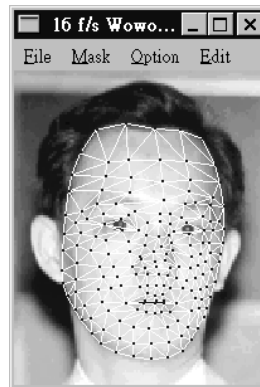
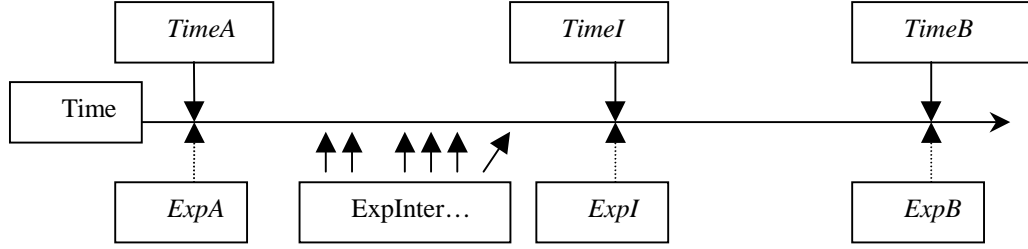


Figure 2: The image is a good example of a neutral face. It is the basis image where the Talk Mask is developed. The origin (0.0,0.0) is set to the upper-left of the image. The image coordinate system sets the lower-right corner as (1.0, 1.0). The base and size of mask is recorded for a specific image.

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. Put the texture mapped mesh model and the background together, the scene now looks just like the original static neutral image.



$$\alpha = (TimeB - TimeI) / (TimeB - TimeA)$$

$$\beta = (TimeI - TimeA) / (TimeB - TimeA)$$

$$ExpressionI = \alpha \times ExpressionA + \beta \times ExpressionB$$

Figure 3: The ExpInter refers to the facial expressions that have been generated from ExpA and ExpB before ExpI was reached.

The first step of facial animation is to define the key frames. The neutral face without any expressions can be seen as a key frames that contains a neutral facial expression. The key frames of different facial expressions such as horror, happy, sad, and anger can be obtained by adjusting the position of different features on the facial parts. From now on, we manipulate the texture-mapped image to edit specific facial expression based on 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.

In a word, 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.

Assume the *Expression A* is synthesized at *Time A*, and is going to be morphed to *Expression B* at *Time B*. The intermediate facial expressions have to be interpolated to

be smooth. As the system runs into loops of generating the intermediate frames, and the program acquires the right to generate *Expression I* at *Time I*. The *Expression I* would be the linear combination of *Expression A* and *Expression B* with appropriate weightings.

By observing the movement of video sequences taken from a subject, the contour of different facial expression can be recorded and classified. According to the analysis of the following paragraphs, in order to pronounce Chinese syllables, which comprises initials middle and finals, less than 10 basic facial expressions are required.

3 Combining with a Chinese Text-To-Speech System

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. The way for Chinese people to understand a speech is to comprehend the context of the tonal speech by reconstructed the syllables into the phrases (a phrase in Chinese comprises usually two to three syllables), which is the major task for natural Chinese speech recognition.

In order to make the talking head make Chinese speech, we have to devise an efficient way to build the visemes, which are the key frames needed for each syllable. 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.

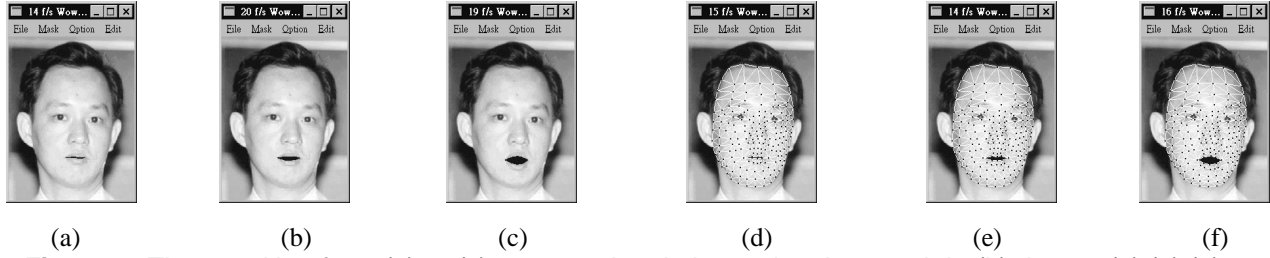


Figure 4: The transition from (a) to (c) at some time is interpolated to result in (b). Image (1),(2),(3) are to show the associated mask mapped on the face.

3.2 Static Key Frames

A key frame will be saved as the label is given. The label is used for viseme script. For instance, the “i” class may contain a key frame labeled “expr1”, the “u” class may contain two key frame labeled “expr2” and “expr3”. Finally the (“i”+“u”) class contains the expressions from “i” and “u” class, which are the “expr1”, “expr2” and the “expr3”.

After defining key frames, we can automatically build the viseme. By defining the duration of between the key frames, the viseme can be animated in real time from one key frame to the other continuously, resulting in the facial animation.

3.3 How To Build Visemes from Key Frames

To pronounce a Mandarin Chinese syllable, one does not always have to make the mouth shape from initials to finals. For plosive sounds such as syllables starting with “bo, po, mo, fo,” one needs to form the specific shape before the utterance.

But for other initials such as “de, te, le, ge, ke, h e” and “gi, qi, xi, zh, ch, sh, r, tz, ts, s,” the mouth shape are largely dissolved into the shape for the finals. During the building process, the initials are included only when the syllable starts with plosive sounds.

4 Speech Synchronazition

By labeling the phoneme, which would be 408 Chinese utterance without tone variation[5], and 1333 Chinese utterance with tone variation, to the visemes, the facial animation can produce all Chinese conversation in real time in theory.

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.

The facial expressions extracted have to synchronize with the syllables pronounced. A precise way to do this is to observing the video sequence of reading process, match the key frames with the sounds. To matching every key frames with every sound manually would be too tedious to be practical. In fact, from the observations, 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.

If a syllable have the time duration of **ad** (see Figure 5), and **n** facial expressions, (**F₁**, **F₂**, ...**F_n**), before time **b**,

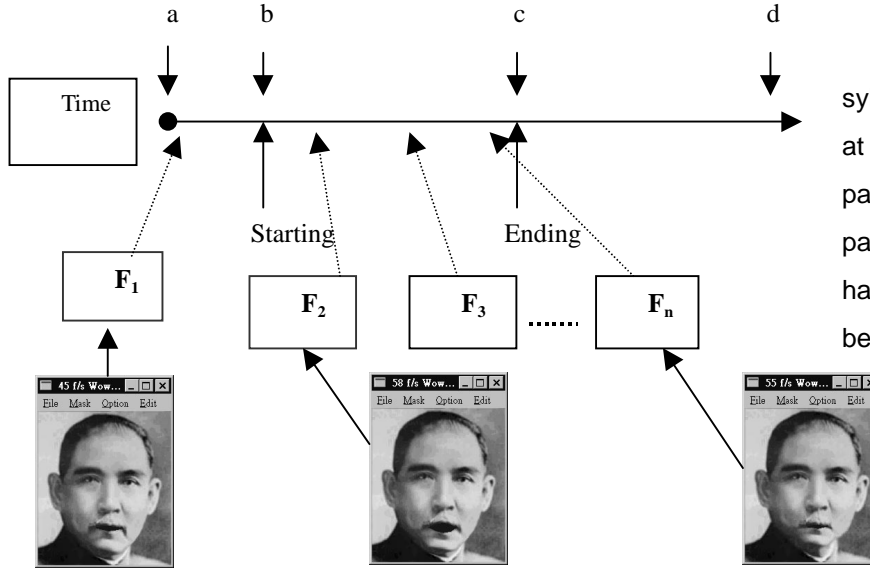


Figure 5: The utterance of a syllables starts from point **a**, and ends at point **d**. Point **b** is the starting parameter, and point **c** is the ending parameter. The facial expressions have to start before **b**, and end before **c**.

the facial expression has to morph to **F₁** 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 single syllable synchronization, before time **a** and after time **d** could be a neutral expression, which means the total sequence would be:

F_{neutral}, F₁, F₂, ..., F_n, F_{neutral}

F_{neutral} is the facial expression before time **a**, and after time **d**.

F₁-F_n viseme begins before time **b** and ends before time **c**.

For this implementation, we set the **F₁-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 **F₁-F_n**, we let them to be evenly divided for smooth transition. Assume the time for a syllable is normalized to be 1.0, the starting 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 (**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, which can be formulated as:

$$\begin{aligned} \mathbf{F}_{\text{transition}} &= rA * \mathbf{F}_n + rB * \mathbf{F}_{\text{neutral}} \\ rB &= (\text{timeFromCtoD} / \text{timeContractionConst}) \\ rA &= (\text{timeContractionConst} - \text{timeFormCtoD}) / \text{timeContractionConst}; \end{aligned}$$

In the implementation, the **timeContractionConst** is set to equal to the **timeSyllable**.

In this mechanism, to produce the current syllable, the **F₁** has to be morphed from a **F_{transition}** from time **a** to time **b**, and be morphed from **F_n** to **F_{transition}** for the next words.

For method (b), the **F₁** is morphed from the last frame of last syllable. Next, the **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 Implementation

The system 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.

Data Manager: Manage the static facial expression and obtain sound elements from Chinese speeches. The facial expression pool and phoneme pool are inside this manager. **Input Manager:** accept input of the Chinese sentence in Big-5 encoded mode. You can also copy and paste part of sentence from another document.

Parsing Manager: parse the input stream. **Syllable Manager:** transform the Chinese sentences into individual Chinese words, and analyze the syllables. Next, it fetches the sound element from phoneme pool, and dissects a syllable into initials and finals. **Viseme Manager:** fetch the facial expressions from the dissected initials and finals and assign them to marked time stamps according to the duration of a given syllable. **Animation Manager:** manage to synchronization the given viseme with the phoneme and produce continuous animation from word to word synchronously. **Output Manager:** Play the sound and facial expressions. The talking head can now pronounce a full sentence.

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 200x300, in true color mode. The performance can reach up to 60 frames/second if the machine is equipped with a hardware accelerator for OpenGL, such as WinFast L2200 (LeadTek Inc.) Image Talk is set to speak three to four words a second. For current implementation, the collected sound elements are restricted inside a 400-millisecond time frame, and the playing sequence can range from 300 to 400 milliseconds for a syllable.

5.2 Generalized Mask

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 the features of a new face. The facial expressions are recorded as deviations from the standard mask, and resized according to the size of the generic mask if the mask was to be matched to a new face. A user can change the image and interactively adjust the mask to match the feature of a new face in real time. The process is done within seconds if one wants to get an approximate match. 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 (see Figure 7) to locally adjust the minor area inside the bounding box. For mapping the

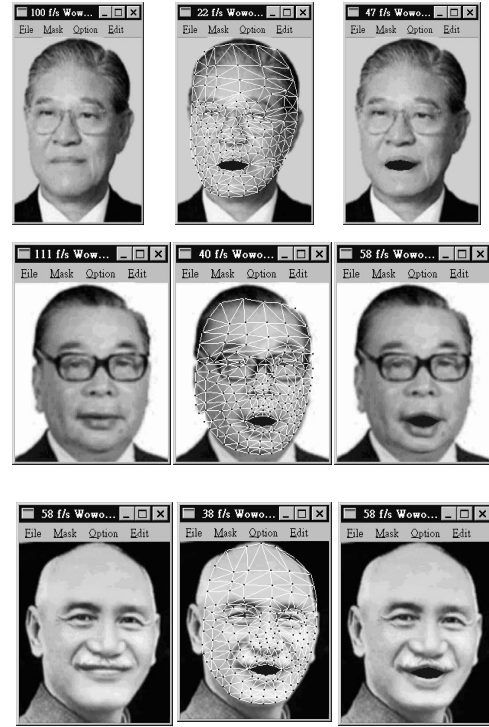


Figure 6: The most left column is the original images downloaded from the Web site of “<http://www.oop.gov.tw>” They were applied with the generic Talk Mask, and are pronouncing the Pinyin symbol “a”. The middle columns are the images with the wired mask. The Talk Mask could be adjusted globally by adjusting the bounding box, or by internal force applied to vertices.

eyes and lips more correctly, a user may takes less than one minute to adjust the mask interactively, and test 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 Head Motion

The head motion is implemented by adjusting the mask with linear transform algorithm described in Figure 7 to be moved to the left, right, up and down to make the head simulate a small amount of rotation. The controlled 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 left-right head motion, but not as good as it is in the up-down motion.

5.4 Eye Blinking

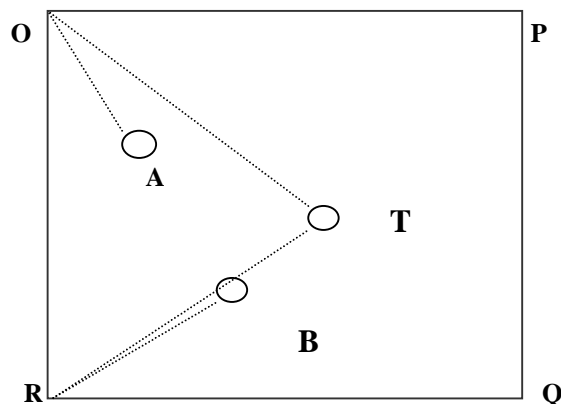


Figure 7: Assume the left is the bounding box of Talk Mask. When drag-with-force function is on: if the user is dragging vertex **T**, and **A** is at the upper left of **T**, Image Talk keeps the ratio of **OA** and **OT** to be of the same while **T** is being manipulated. For **B**, which is at the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

Semi-periodic eye blinking are added to make the face looks like live human. Image Talk does not really process the eyes right now. The 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.

5.5 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 Conclusions

Image Talk produces a lifelike talking head. When applied to other images, it can even make a deceased man talk. The eye blinking, head motion, and lip-sync combined with Mandarin Chinese text-to-speech make the system quite entertaining.

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 left-right rotation, but not quite well in the up-down rotation. A 3D head model

can be applied to solve the problem.

Future Work: Image Talk does not really process the eye part right now. The blinking eyelid is done by pulling the upper facial area of the eyes. The eyeballs can be single out to make synthetic ones, in order to change the viewing direction of the eyes. Similarly, teeth and tongue should be displayed. The Chinese Text-to-speech system can be easily replaced by other ones. The fidelity of the system would definitely be doubled if this interface were integrated with a commercial text-to-speech system.

Reference

- [1] K. Aizawa, Thomas S. Huang. "Model-based image coding: advanced video coding techniques for very low bit-rate applications." *Proceedings of the IEEE*. 83(2), pp.259-271, February 1995.
- [2] Thaddeus Beier, Shawn Neely. *Feature-based image metamorphosis*. *Computer Graphics*, 26(2), pp.35-42, 1992. ISSN 0097-8930
- [3] Christoph Bregler, Michele Covell, Malcolm Slaney. "Video Rewrite: Driving Visual Speech with Audio." *Computer Graphics Proceeding (SIGGRAPH 97)*, pp.353-360, 1997.
- [4] 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.
- [5] Lin-Shan Lee, Chiu-Yu Tseng, Ming Ouhyoung, "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.
- [6] Yuencheng Lee, Demetri Terzopoulos, Keith Waters. "Realistic Modeling for Facial Animation." *Computer Graphics Proceedings (SIGGRAPH 95)*, pp.55-62, 1995.
- [7] P. Litwinowicz, L. Williams. "Animating images with drawing." *SIGGRAPH94*, Orlando, FL, pp.409-412, 1994. ISBN 0-89791-667-0
- [8] Massaki Oka, Kyoya Tsutsui, Akio Ohba, Yoshitaka Kurauchi, Takashi Tago. "Real-time manipulation of texture-mapped surfaces." *ACM Computer Graphics (SIGGRAPH 87)*, 21(4), pp.181-188, 1987.
- [9] Keith Waters. "A Muscle Model for Animating three-dimensional Facial Expression." *ACM Computer Graphics (SIGGRAPH 87)*, 21(4), pp.17-24, 1987.
- [10] Lance Williams. "Performance-Driven Facial Animation." *ACM Computer Graphics (SIGGRAPH 90)*, pp.235-242, 1990.

- [11] F. Parke. Computer generated animation of faces. In ACM National Conference, pages 451-457. ACM, 1972.
- [12] F. Parke. Parameterized models for facial animation. IEEE Computer Graphics and Applications, 2(9):61-68, November 1982.
- [13] Frederic Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski and David H. Salesin. Synthesizing Realistic Facial Expressions from Photographs, SIGGRAPH 98, 1998