Synthesis of Facial Images with Lip Motion from Several Real Views

Lei Gao† Yasuhiro Mukaigawa‡ Yuichi Ohta†

† Institute of Information Sciences and Electronics
University of Tsukuba
Tsukuba, Ibaraki, 305-8573 Japan
{gao, ohta}@image.is.tsukuba.ac.jp

‡ Department of Information Technology, Faculty of Engineering
Okayama University
Tsushima-naka 3-1-1, Okayama, 700-8530 Japan
mukaigaw@chino.it.okayama-u.ac.jp

Abstract

The synthesis of facial images by computer graphics is very important for many applications such as human interface and visual entertainment. The lip motion is an essential factor in synthesizing the image sequence of conversation. In this paper, we propose a new method for synthesizing facial images with lip motion.

The key feature of our system is that it does not need any models of lip motion. Arbitrary lip shapes are expressed by the combination of several real views. By using several images with basic lip shapes, facial image sequence with lip motion in conversation can be generated well by their linear combination.

1. Introduction

In recent years there has been considerable interest in the human-computer interface. The applications, such as virtual actors and teleconferencing, require the techniques of synthesizing realistic facial images which express the various nuances of facial motion [7]. To generate an image sequence of speakers in teleconferencing, it is a very important factor to synthesize natural lip motion in conversation.

In order to describe the lip motions in speech, lip models are often used and lip shapes are parameterized [9]. Many model-based approaches try to recognize the lip shape from a video sequence. For instance, Cootes et al. [4] developed 'Active Shape Models' and described how the models can be used in local image search. A kind of statistical model was used in their work to model the variation of the face shape and it was fit to a new image. Bregler et al. [3] built a parameterized model of the complex "space of lip configuration" by using machine learning. Their system is given a collection of training images which were used to construct the model for the lip recognition. Of course, the result of visual speech recognitions can be used for image synthesis. However, the accurate modeling of lip shape in visual speech recognitions is often troublesome and it is difficult to generate a natural lip motion. We can say that both of the modeling of lip motion and the visual speech recognition are not necessary for the lip image synthesis.

In this paper, we propose a new method that generates natural conversational image sequence with lip motions by combination of several real views without any modeling. Mukaigawa et al. [5] [6] proposed a method for synthesizing facial views with arbitrary expressions from some representative expressions. We apply this principle to synthesizing lip motions and show that various conversational image sequences can be generated from the linear combination of several vowel lip shapes.

There are some related works. For instance, Bregler et al. [2] proposed Video-Rewrite which prepares the lip image sequences for each pronunciation, and switches them by interpolation. Our method has an advantage that the image generator needs only a small number of lip images, because the lip images for each pronunciation can be synthesized from basic lip images. Bascle et al. [1] presented an approach which tracks head movement and facial expression and makes facial animations drawn by lines. Our method can synthesize real facial images by using texture blending.

2. Expression of Lip Shapes

In this section, we show that an arbitrary lip shape can be synthesized by combining some basic lip shapes.

2.1. Determination of Feature Points

Let $B_1, \ldots, B_m$ be a set of face images, which we call base images, with different lip shape. We assume that n feature points are located on the face, and that the correspondences of all feature points among all base images have been known. We try to synthesize a new facial image with an arbitrary lip shape from these base images.

Assume that the coordinate of i-th $(1 \le i \le n)$ feature point on B_j $(1 \le j \le m)$ is (x_i^j, y_i^j). The X- and Y-coordinates of all feature points can be expressed as vector $\boldsymbol{x}^j$ and $\boldsymbol{y}^j$ as shown in Equation (1).

$$\begin{aligned} \boldsymbol{x}^j &= [x_1^j, x_2^j, \ldots, x_n^j] \\ \boldsymbol{y}^j &= [y_1^j, y_2^j, \ldots, y_n^j] \end{aligned} \tag{1}$$

Let $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ be vectors which indicate the X- and Y-coordinates of all feature points on a new view with an arbitrary lip shape. If the base images include a sufficient variety of lip shapes, $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ can be approximated by the linear combination of $\boldsymbol{x}^j$ and $\boldsymbol{y}^j$ as shown in Equation (2).

$$\begin{aligned} \hat{\boldsymbol{x}} &= \textstyle\sum_{j=1}^m c_j \boldsymbol{x}^j \\ \hat{\boldsymbol{y}} &= \textstyle\sum_{j=1}^m c_j \boldsymbol{y}^j \end{aligned} \tag{2}$$

Usually the coordinate vectors $\boldsymbol{x}^j$ and $\boldsymbol{y}^j$ are not linearly independent and can not be used as bases in the vector space. This may cause instability of the coefficients for the linear combination. So here we perform principal component analysis on the coordinate vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$ and $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^m$ of the set of base images and get the principal components $\boldsymbol{P}^1, \ldots, \boldsymbol{P}^m$. By using those principal components as bases, the effect of noise can also be reduced. With the top k principal components, we can rewrite Equation (2) as the following:

$$\begin{aligned} \hat{\boldsymbol{x}} &= \textstyle\sum_{j=1}^k C_j^x \boldsymbol{P}^j \\ \hat{\boldsymbol{y}} &= \textstyle\sum_{j=1}^k C_j^y \boldsymbol{P}^j \end{aligned} \tag{3}$$

2.2. Texture of Face Images

In order to synthesize a facial image after the coordinates of feature points are calculated, triangular patches whose vertices are those feature points are created. Textures taken from the base images are mapped to the triangular patches.

However, if we take all textures from one base image, the synthesized image will be warped unnaturally as the lip shape changes. This undesirable warping is caused by a drastic deformation of texture. In order to make full use of the base images, we use texture blending. In our method, facial images are synthesized by mapping the blended texture taken from multiple base images. If the lip shape on the synthesized facial image is similar to that on one base image, the weight of texture of this base image is set larger, otherwise it is set smaller.

The texture taken from each base image B_j is blended with a weight parameter w_j, with $\sum w_j = 1$. Here, w_j is decided to be inverse of the distance between this base image B_j and the synthesized image $\hat{B}$ in the principal component space as illustrated in Figure 1.

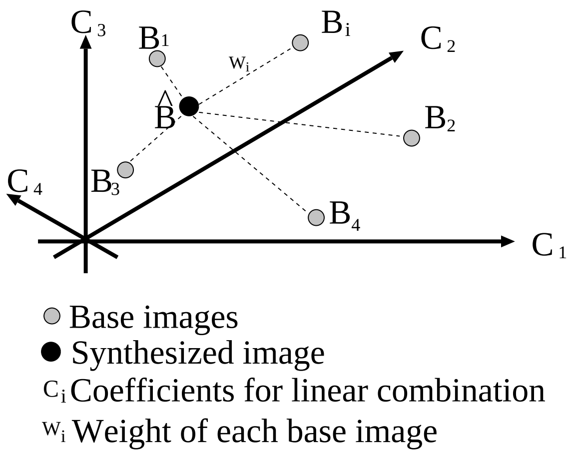

Figure 1. Euclid distance in the principal component space

3. Synthesis of Lip Motions

In the sections above, we show that facial images with arbitrary lip shape can be synthesized according to the coefficients for the linear combination of several basic lip shapes. In this section, we discuss about how to determine these coefficients to generate a conversational image sequence.

In order to determine the sequence of coefficients corresponding to a conversational image sequence, we use a sample image sequence. In this way, image sequence that has the same lip motion as the sample sequence can be synthesized by the combination of several base images, as shown in Figure 2.

Given several feature points on the sample images, whose number can be less than that of feature points on the base images, the coefficients can be calculated by Equation(3). If the number of the feature points on the sample images is more than that of the base images, a least square method is used.

If the actor in base images is the same person as that in the sample sequence, we can re-synthesize the same image sequence as the sample one by specifying the coefficients of each. This is useful for some applications such as image compression and telephotography.

If the actor in base images is a different person from that in sample sequence, we can regenerate a conversational sequence of the actor with the same lip motion as that of the sample person. This is an important technique for sound dubbing or virtual actors.

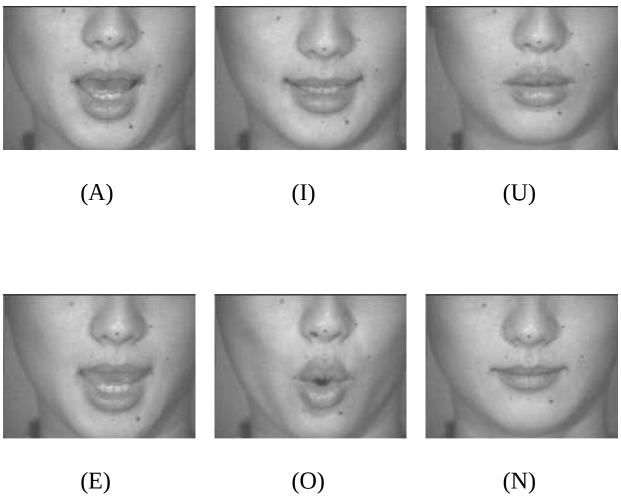

Figure 3. Base images used for experiment A

4. Experimental Results

Here we show two experiments on the synthesis of face images with lip motion from several real views. One is performed by using the same person's base images as the sample images. The other is carried out by using base images of different person from the sample.

In these experiments, the coordinates of the all feature points are given manually by referring to the marks drawn around the lip. This is the simplest approach for tracking the lip motion.

4.1. Experiment A: Using the same person's sample sequence

In general we do not know what kind of lip shape should be used as base images to synthesize the images with arbitrary lip motion. Here we present two methods to select lip shapes used for the base images.

According to the lip-reading system, the lip shapes of the five vowels A, I, U, E, O in Japanese and the lip shape of closed mouth considered as phoneme N, should be regarded as basic lip shapes [8]. Thus, first of all, we take the images with these six lip shapes, which are used as base images as shown in Figure 3.

On each base image, 32 feature points are detected manually. Then the triangular patches which have 56 facets are created as shown in Figure 4.

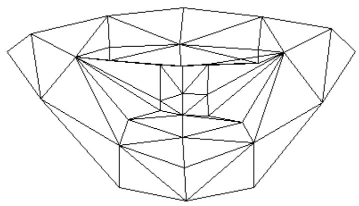

Figure 4. Triangular patches

Figure 5 shows the sample image sequence: "konnichiwa", which means "hello" in Japanese. On each sample image, only 10 feature points are detected. The coefficients of linear combination are calculated according to these detected feature points, and the facial image sequence is synthesized as shown in Figure 6. We can see that lip shapes in conversation can actually be synthesized from some other lip shapes. Although the synthesized images are not equal to the sample images exactly, we can 'read' what the person is talking about from the lip motions.

Another method for determining a set of base images is that we use principal component analysis on some conversational image sequences, and select several top

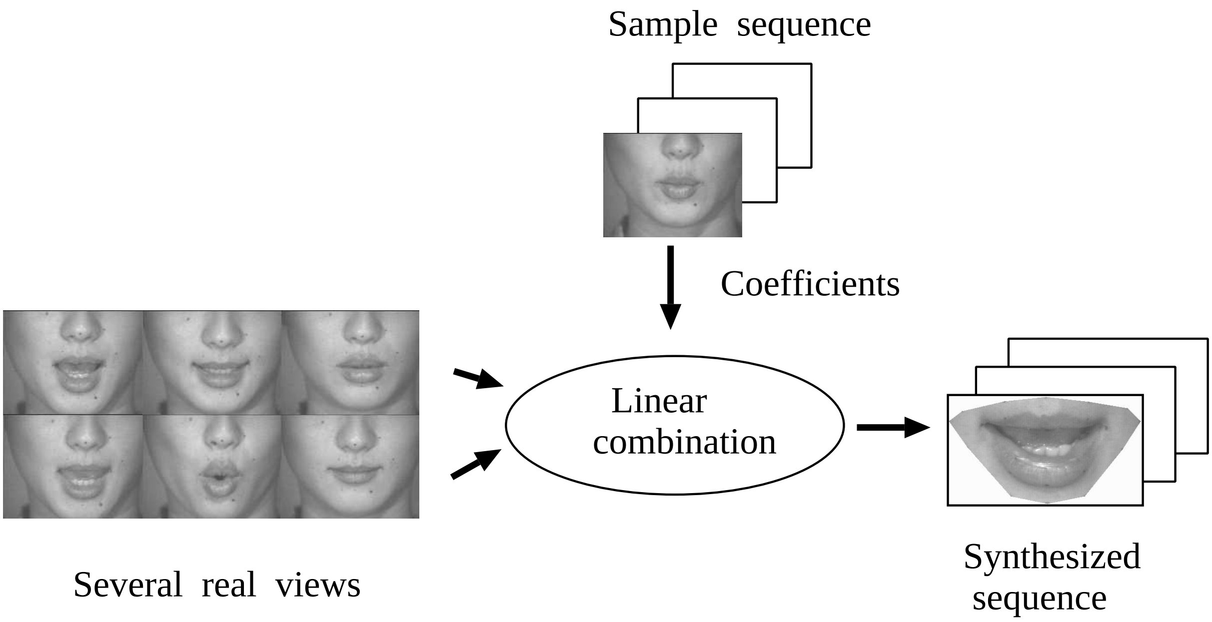

Figure 2. Synthesis of lip motions

Figure 5. Sample image sequence: "konnichiwa"

principal components as the base vectors. Since the eigen values corresponding to those top principal components are bigger than the others, we can calculate the coordinates of feature points with the combination of only a few base vectors. As an example in our experiment, we carried out principal component analysis on a conversation image sequence: "ohayougozaimasu", which means "good morning" in Japanese. We used the top four principal components as the base vectors. Table 1 shows the result of principal component analysis. Figure 7 shows the synthesized images from the obtained base vectors and selected representative textures.

Table 1. Result of principal component analysis

PRINCIPAL COMPONENTS	EIGEN VALUE	PROPORT	CUMULATE
1	2499.9	60.88%	60.88%
2	1200.9	29.25%	90.13%
3	108.1	2.63%	92.76%
4	87.5	2.13%	94.89%
5	60.8	1.48%	96.37%

4.2. Experiment B: Using different person's sample sequence

It should be more interesting and useful in human interface and entertainment visual system to create a new speaker's conversational image sequence. We present here an experimental result of synthesizing different person's image sequence from sample images. Figure 8 shows the base images with lip shapes of A, I, U, E, O and N in Japanese. 52 feature points were marked on the face and the triangular patch model has 86 facets.

Figure 6. Synthesized image sequence from 6 base images

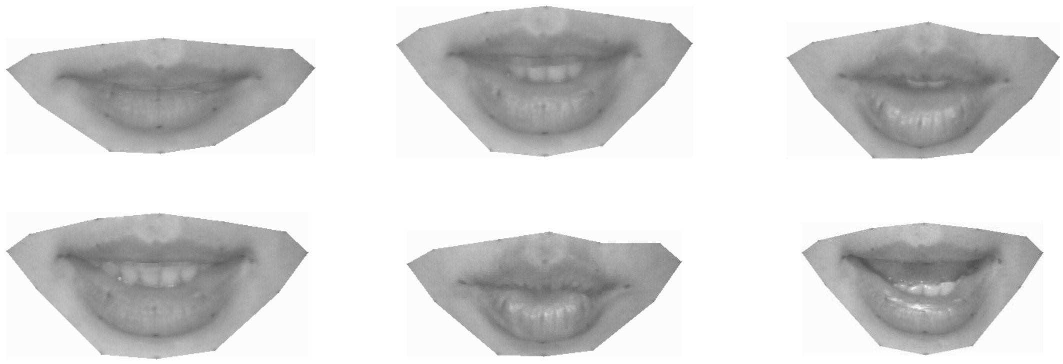

Figure 7. Synthesized images from principal components

The sample image sequence is the same conversational sequence as that used in Experiment A.

In this case, new conversational views generally can not be well synthesized, since different person has different mouth shape and the same feature points can not be precisely matched. So we use the same coefficients in synthesizing images as that used in Experiment A. Figure 9 shows some images from the synthesized conversational sequence. Although there are some unnatural parts in synthesized images compared with the sample views, we still can 'read' what the person is speaking from them.

5. Discussion

We have proposed a method for the synthesis of conversational image sequence with lip motion. The experimental results show that conversational images with lip motion can be well synthesized by using linear combination of several base images without the modeling of lip motions.

It can be expected that more natural conversational views can be well synthesized if more base images are used. There still remains, however, some questions such as how to evaluate the synthesized images. Although some subjective evaluations can be done, it can not totally depend upon if one can accurately 'read' out what the conversational image sequence 'says'. Some objective evaluation methods can also be used such as calculating the errors of 2D coordinates of feature points between the synthesized views and real conversational image sequence.

The difficult task of automatically tracking lip motion without invasive markers also remains an active area of research.

References

[1] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. *Proc. of Int. Conference on Computer Vision (ICCV'98)*, pages 323–328, 1993.

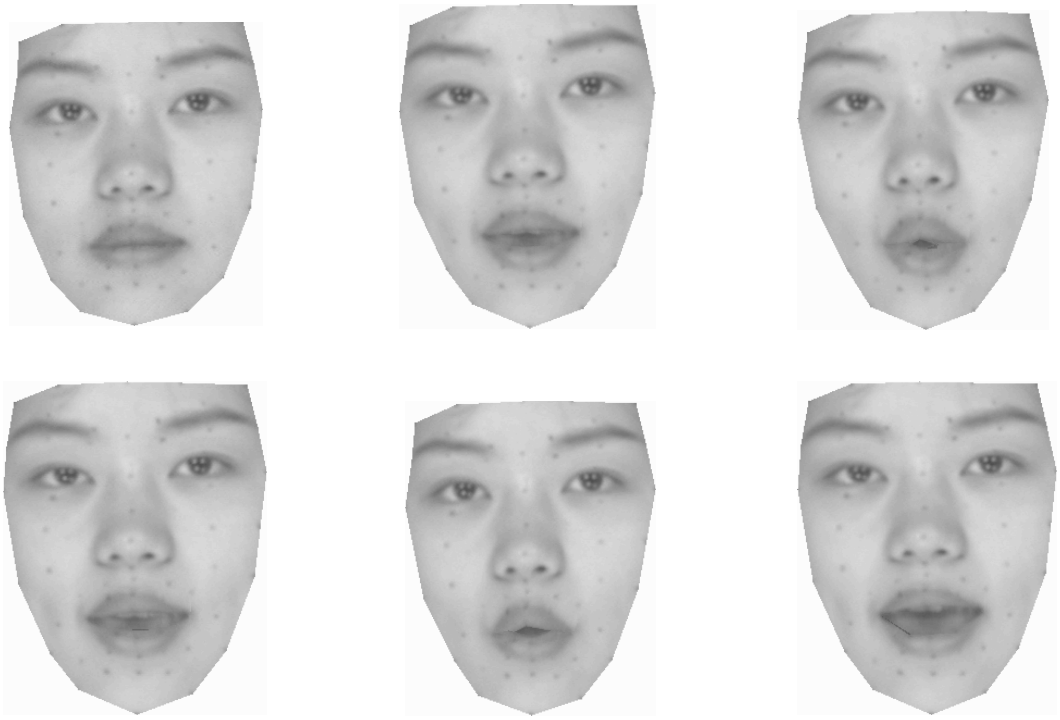

Figure 9. Synthesized image sequence in experiment B

[2] C. Bregler, M. Covell, and M. Alaney. Video rewrite: Driving visual speech with audio. *Computer Graphics Proceedings, Annual Conference Series*, pages 353–360, 1997.

[3] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. *Proc. of Int. Conference on Computer Vision (ICCV'95)*, pages 494–499, 1995.

[4] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. *Proc. of Int. Conference on Computer Vision (ICCV'93)*, pages 242–246, 1993.

[5] Y. Mukaigawa, Y. Nakamura, and Y. Ohta. Synthesis of arbitrarily oriented face views from two images. *Proc. of Asian Conference on Computer Vision (ACCV'95)*, 3:718–722, 1995.

[6] Y. Mukaigawa, Y. Nakamura, and Y. Ohta. Face synthesis with arbitrary pose and expression from several images - an integration of image-based and model-based approach -. *Proc. of Asian Conference on Computer Vision (ACCV'98)*, 1:680–687, 1998.

[7] F. I. Parke and K. Waters. *Computer Facial Animation.* A K Peters, Ltd., Wellesley, Massachusetts, 1996.

[8] H. Sera, S. Morishima, and D. Terzopoulos. Synthesis and analysis of speech animation with physics-based muscle model. *Proceedinds of the 2nd Symposium on Intelligent Information Media*, pages 83–90, 1996.

[9] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(4):235–242, 1990.

(A) (I) (U)

(E) (O) (N)

Figure 8. Base images used for experiment B