THE USE OF DYNAMIC DEFORMABLE TEMPLATES FOR LIP TRACKING IN AN AUDIO-VISUAL CORPUS WITH LARGE VARIATIONS IN HEAD POSE, FACE ILLUMINATION AND LIP SHAPES

Zhiyong Wu[1,2], Jiying Wu[1] and Helen M. Meng[1,2]

[1] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR
[2] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055

ABSTRACT

This paper describes an approach for lip tracking using dynamic deformable templates. The objective is to track lip parameters from an audio-visual corpus recording a voice talent who is reading text prompts in a natural and expressive way. The corpus presents challenges to the conventional method of lip tracking with deformable templates. This is because natural and expressive speech includes relatively large motions of the head and the lips. The head motions lead to changes in the illumination of the face region and changes in the observed lip shape. In addition, emphatic pronunciations lead to large changes in the lip shape. Video frames that are affected by face illumination changes present additional difficulty in locating the mouth region (i.e. region of interest, ROI). Video frames that are affected by changes in lip shapes present additional deviations from the lip templates and hence lower tracking accuracies. Our proposed method incorporates "dynamicity" in the deformable templates to render them adaptive to changes in head pose, face illumination and lip shapes. Experiments show that dynamic deformable templates consistently outperform the conventional deformable templates in lip tracking.

Index Terms— lip tracking, dynamic deformable template, text-to-audio-visual-speech (TTAVS) synthesis

1. INTRODUCTION

Human speech is bimodal in nature [1]. There is much to be gained by leveraging the complementary and redundant relations between the audio and visual modalities to enhance human-computer speech communication, such as audio visual speech recognition [2], audio visual speaker verification [3], text-to-audio-visual-speech (TTAVS) synthesis [4], etc.

The objective of our research is to understand the temporal relations between the audio and visual modalities for TTAVS synthesis in Chinese Cantonese and Putonghua (Mandarin). For this purpose, we need to perform lip tracking on an existing audio-visual corpus, the CU-TTAVS corpus, to extract visual features. Unlike other audio-visual corpora [5], CU-TTAVS corpus contains natural head and lip motions of a speaker reading a large set of text prompts. Hence it contains large variations in head pose and lip shapes, which present challenges for lip tracking as will be described later. The head motions lead to variations in the illumination of the face region and the observed lip shapes.

There are many methods proposed for lip tracking, such as snakes [6], active appearance models (AAM) [7], active shape models (ASM) [8], deformable templates (DT) [9], etc. The need for many free parameters is the drawback of snakes in real application. AAM/ASM requires a large number of training data for building the model. DT uses predefined templates controlled by an energy function to approximate lip contours, and can detect lip contours much faster than snakes. DT generally requires less training data than AAM/ASM. Hence we believe that DT is suitable for real applications and is adopted in this work. However, the presence of large variations in head pose causes distortions in the observed lip shapes and illumination differences in the face region, which present difficulties to lip tracking. Emphatic lip motions captured from expressive speeches also present similar difficulties.

We present a novel dynamic deformable template method for lip tracking in the CU-TTAVS corpus. Unlike conventional DT, the proposed method can adaptively search for "regions of interest" (ROIs) based on the illumination conditions and head pose of current video frame.

Table 1. *Statistics for the CU-TTAVS corpus (Unit: in utterance).*

450 Cantonese utterances			
Low-level lighting conditions	Varying lip shapes		Varying head poses
	Normal	Tightly closed lips	
9	128	219	94
700 Putonghua utterances			
Low-level lighting conditions	Varying lip shapes		Varying head poses
	Normal	Tightly closed lips	
77	80	80	463

2. ANALYSIS OF THE CU-TTAVS CORPUS

The CU-TTAVS corpus aims to represent different variations including head pose, lip shapes and lighting conditions in real situation of speech communication. There are in total 450 Cantonese and 700 Putonghua utterances. Among them, 86 utterances are collected with low-level lighting

conditions; 567 utterances are recorded with large head motions; the others have the standard frontal face with different lip shapes (i.e. normal or tightly closed lips). Detailed statistics of the corpus are shown in Table 1.

3. BACKGROUND ON LIP TRACKING

3.1. Parameters in lip tracking

We adopt the MPEG-4 Facial Animation (FA) standard for facial parameterization [10]. Since our work on lip tracking is used for TTAVS, we mainly focus on the feature points in nose and lip regions, especially the 6 points for lip tracking as illustrated by N_1 to N_6 in Figure 1:

- 2 points for nostrils: these are labeled as N_1 and N_2 in Figure 1. W_1 denotes the width between two nostrils.
- 4 points for lips: these include the left lip corner (N_3), the right lip corner (N_4), the midpoint of upper lip (N_5) and the lowest point of lower lip (N_6). H denotes the height of the lips, obtained from N_5 and N_6.

As will be elaborated later, some of these points help delineate three "regions of interest" (ROIs) (see Figure 1) for tracking the two nostril points and the lip points respectively.

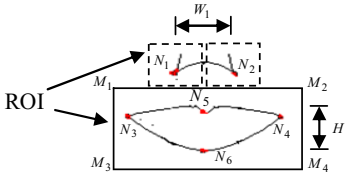


Figure 1. *Feature points and regions of interest for lip tracking.*

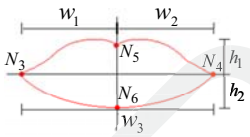


Figure 2. *Basic template for lip tracking.*

3.2. Lip tracking by deformable template method

The deformable template method proposed in [9] has been widely used for lip tracking. As illustrated in Figure 2, three quartics (two for upper lips, one for lower lip) comprise a lip template which is adjusted to approximate the lip contours. Each quartic takes the form:

$$y = h(1-\frac{x^2}{w^2}) + 4q(\frac{x^4}{w^4} - \frac{x^2}{w^2}) \qquad (1)$$

where w and h are the width and height of lip contour curve (see Figure 2). q(=1) is the parameter controlling the curve shape. The positions of lip points are hand-defined in the first video frame to generate the initial parameters (i.e. w, h) for template. Then the template deviations are controlled by a temporal penalty energy function in the subsequent frames to extract the lip points in each frame. The energy function is defined to evaluate whether or not a point belongs to the lip region (lip ROI) based on color and edge information.

The 2 nostril ROIs are utilized in the determination of the lip ROI. These 2 ROIs are derived from the nostril points of the previous frame N_1 (x_{N1}, y_{N1}) and N_2 (x_{N2}, y_{N2}) as shown in Figure 1. Variable L is defined as the half distance between the two nostrils (Equation 2). Thereafter, nostril points of current frame are searched within the two ROIs (x_{N1}–$L \le x \le x_{N1}+L$, y_{N1}–$L \le y \le y_{N1}+L$) and (x_{N2}–$L \le x \le x_{N2}+L$, y_{N2}–$L \le y \le y_{N2}+L$), and detected as the central point of the black area after binarizing the image in ROIs.

$$L = \frac{\sqrt{(x_{N_1} - x_{N_2})^2 + (y_{N_1} - y_{N_2})^2} + 1}{2} \qquad (2)$$

The lip ROI is then derived from nostril points in current frame and lip points in previous frame. The coordinates of the diagonal corners M_1 (x_{M1}, y_{M1}) and M_4 (x_{M4}, y_{M4}) of the lip ROI as shown in Figure 1 are defined as:

$$x_{M_1} = x_{N_1} - W_1 * 2 \qquad x_{M_4} = x_{N_2} + W_1 * 2$$
$$y_{M_1} = (y_{N_1} + y_{N_2})/2 + W/2 \qquad y_{M_4} = (y_{N_5} + y_{N_6})/2 + H \qquad (3)$$

4. LIMITATIONS OF THE CONVENTIONAL DEFORMABLE TEMPLATE (DT) METHOD

The size and position of the ROIs are crucial for lip tracking with the deformable template method. Actual ROIs vary with large variations in head pose, face illumination and lip shapes. This may affect lip tracking performance drastically.

4.1. Limitations of DT for low-level lighting conditions

The first limitation of the conventional DT is that it can not handle differences in color pixel distributions, especially under low-level lighting conditions. Figures 3(a) and (b) illustrate the frontal face samples collected under high-level and low-level lighting conditions respectively. The shadow around nose in Figure 3(b) affects the binarizing process of the facial image and will lead to incorrect nostril detection result (see Figure 5(a) for an illustration). As a result, the ROI for lip region will be falsely located according to incorrect nostril positions. This may eventually lead to wrong lip points tracking results.

(a) high illumination condition (b) low illumination condition

Figure 3. *Face samples under different lighting conditions.*

(a) frontal face sample (b) face sample with tilted pose

Figure 4. *Face samples with different head poses.*

4.2. Limitations of DT on non-frontal facial images

The second limitation is related to head pose variations. In conventional DT, the lip ROI is set to be large enough to contain all possible lip points. However, this may lead to the inclusion of non-facial features (e.g. hair, collar) in the

ROI on non-frontal facial images. Figure 4 shows the sample of such situation. In Figure 4(b), the speaker lowers her head and turns left; the lip points could not be precisely located due to the influence of the microphone and collar.

4.3. Limitations due to large variations in lip shapes

The last limitation of the conventional DT is that it can not model the lip shape with large deviations from the template. Conventional DT assumes the lip shape to be 4th order curve using 3 quartics. But in CU-TTAVS corpus, the speaker often closes her lips tightly (Figure 9), where the 3 quartics degrade into just one quadratic or even a line in such circumstances, leading to incorrect lip tracking results.

5. THE DYNAMIC DEFORMABLE TEMPLATE METHOD

A dynamic deformable template method is proposed for lip tracking in the CU-TTAVS corpus with large variations in head pose, face illumination and lip shapes.

5.1. Nostril searching according to lighting conditions

The first dynamic adaptation is for nostril search under variable illumination conditions. As illustrated in Figure 5(a), the nostrils might be falsely located under low illumination condition because of the difficulty in discriminating color pixels of the nostrils from the background skin color due to the shadow around nose. To solve this problem, the ROI for nostrils is redefined so as it can dynamically fit different conditions of illumination:

$$L = \frac{\sqrt{(x_{N_1} - x_{N_2})^2 + (y_{N_1} - y_{N_2})^2} + 1}{\alpha}, \quad \alpha \geq 2 \tag{4}$$

where α is dynamically determined by lighting condition. The illumination is lower, α is larger.

Several dynamic values of α can be considered with respect to complex illumination conditions. In this paper, the initial value of α is empirically set to be 4. The nostrils are searched within a relatively small ROI based on the L value computed with α=4; and thus the fewer pixels with skin color would affect the nostril detection result. If the nostril can not be found, α is dynamically changed to 2 to define a bigger ROI for searching. Figure 5(b) shows the nostrils searching result of the dynamic method, where right nostril is correctly located.

(a) result of conventional method (b) result of dynamic method

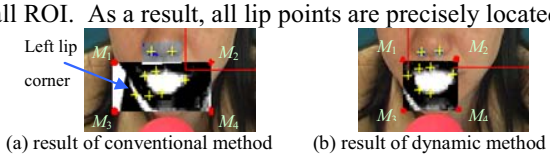
Figure 5. *Nostril searching result under low-level lighting.*

5.2. Lip tracking on non-frontal facial images

The second dynamic adaptation is for lip tracking on facial images with varying head poses. Figure 6(a) shows the lip tracking result of the conventional method, where the left lip corner has been falsely placed on the collar. It is caused by the influence of non-facial features (i.e. collar) contained in the large ROI. To solve this problem, the width of the lip ROI is redefined to produce the small ROI for lip region:

$$x_{M_1} = x_{N_1} - W_1, \quad x_{M4} = x_{N2} + W_1 \tag{5}$$

Figure 6(b) shows the lip tracking result using the refined parameters, where only facial features are contained in the small ROI. As a result, all lip points are precisely located.



(a) result of conventional method (b) result of dynamic method

Figure 6. *Lip tracking result on non-frontal facial image.*

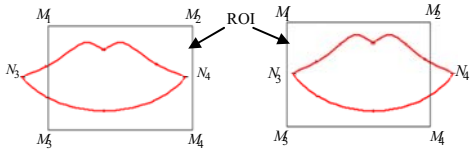


Figure 7. *Lip corner points exceeding the ROI.*

However, the lip corner points may exceed the small ROI as illustrated in Figure 7. Another dynamic method is proposed to fix this problem, as described in Table 2. The size of the lip ROI is dynamically determined by the detection result of lip corner points. When the left/right lip corner is detected on the left/right edge of the ROI (i.e. x_{N3}=x_{M1} or x_{N4}=x_{M4}), it indicates that lip corner points may exceed the small ROI and the size of ROI should be enlarged dynamically. With this dynamic method, all the lip points can be precisely tracked in CU-TTAVS corpus.

Table 2. *Dynamic algorithm for position and size of the lip ROI.*

x_{M1}=x_{N1}-W_1, x_{M4}=x_{N2}+W_1;
Lip tracking using deformable template;
if (x_{N3}=x_{M1}) or (x_{N4}=x_{M4})
  Re-calculate the value of lip ROI:
    x_{M1}=x_{N1}-W_1*2, x_{M4}=x_{N2}+W_1*2;
  Lip tracking using deformable template;
end if;
Output the lip tracking results.

5.3. Lip tracking on facial image with varying lip shapes

The last dynamic adaptation of our method is for tracking lip contours with varying lip shapes. When the lip shape degrades into a curve or a line because the speaker's lips are tightly closed, a threshold H^* for the mouth height H (see Figure 1) is experimentally defined as:

$$H^* = 0.1 * \frac{\sum_{i=1}^{n} H_i}{n}, \quad n = 557 \tag{6}$$

557 utterances in CU-TTAVS corpus with different head poses and lip shapes have been used as the training set to calculate the threshold. The H_i in Equation (6) is the average mouth height in utterance i. While tracking the lip contours for the image frames of a new utterance, if H is smaller than H^*, the lip shape is assumed to be a line; and a parabolic curve is used to approximate the lip contour:

$$y = h'(1 - \frac{x^2}{w'^2}) \tag{7}$$

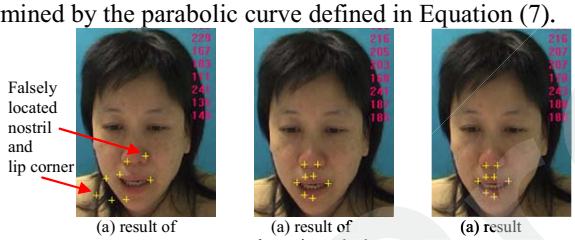
where w' and h' are the width and height of the parabolic curve. The lip corner points are determined the same as

before; the other lip points are not generated by the template.

6. EXPERIMENTS

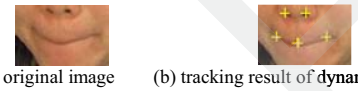
As an illustration, Figure 8 shows the lip tracking results of the non-frontal facial image under low-level lighting condition. Figure 8(a) gives the result of the conventional DT. As illustrated by the arrows, the right nostril is incorrectly located due to the low illumination. The left lip corner is falsely located below the face because the lip ROI is too big and affected by non-facial features (i.e. hair). These distortions are caused by the tilted head pose. Figure 8(b) shows the result of the proposed dynamic DT method. The right nostril point is properly located and the lip feature points are precisely extracted by restricting the ROIs. The manually annotated nostril and lip points are shown in Figure 8(c). This illustrates the superiority of the dynamic DT method.

The lip tracking results on images with varying lip shapes are also evaluated. Figure 9(a) shows an image sample from CU-TTAVS corpus, where the speaker tightly closes her lips. It is impossible to discriminate the upper and lower lips. In this situation, the lip contour is approximated by a parabolic curve as described in section 5.3. The lip tracking result is shown in figure 9(b), where the central point is determined by the parabolic curve defined in Equation (7).



(a) result of conventional method (a) result of dynamic method (a) result manually annotated

Figure 8. *Lip tracking result under low-level lighting condition and on non-frontal facial image.*



(a) original image (b) tracking result of dynamic method

Figure 9. *Lip tracking result for the tightly closed lips.*

We use the mean square error (MSE) to evaluate the performance of the lip tracking methods. The automatically tracked feature points are compared with the manually annotated feature points.

$$MSE_i = \frac{1}{20}\sum_{j=1}^{20}\sqrt{(x_j^{i'} - x_j^i)^2 + (y_j^{i'} - y_j^i)^2}$$
$$MSE = \sum_{i=1}^{4} MSE_i \qquad (8)$$

where $x_j^{i'}$, $y_j^{i'}$ are the coordinates of lip point i (i.e. in Figure 1) on image frame j generated by the DT method; x_j^i, y_j^i are the corresponding coordinates annotated manually. The MSE values for conventional and dynamic DT methods are calculated on 20 frames with above various conditions. As can been seen from the results shown in Table 3, MSE values of the newly proposed dynamic DT method are lower than those of the conventional method under all conditions. This indicates the validity of the dynamic DT method by overcoming the limitations of the conventional DT method.

Table 3. *MSE for different methods under various conditions.*

	Low-level lighting	Varying head poses	Varying lip shapes
Conventional DT method	427.98	40.86	59.29
Dynamic DT method	22.12	32.70	31.97

7. CONCLUSIONS

This paper proposes a dynamic deformable template method for lip tracking. The objective is to track lip parameters for TTAVS from an audio-visual corpus recorded in a natural and expressive way. The large head and lip motions in the corpus present challenges to the conventional deformable template. Head motions lead to changes in face illumination and observed lip shape. Emphatic pronunciations lead to large changes in lip shape also. Face illumination changes present additional difficulty in locating the ROIs. Lip shape changes present additional deviations from lip templates. Our proposed method incorporates "dynamicity" in the deformable templates to render them adaptive to changes in head pose, face illumination and lip shapes. Experiments show that dynamic deformable templates outperform the conventional deformable templates in lip tracking.

8. ACKNOWLEDGEMENTS

This work is partially supported by the grant from the Hong Kong SAR Government's Research Grants Council (RGC) Earmarked Grant (CUHK4149/06E).

9. REFERENCES

[1] C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces", Journal on Communications of the Scientific Soc. For Telecommunications, 43: 32-40, 1992.

[2] C. Neti, G. Potamianos., L. Luettin, et al., "Audio visual speech recognition, Final workshop 2000 report", Tech.Rep., Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA, 2000.

[3] J. Luttin, N.A. Thacker, S.W. Beet, "Speaker identification by lipreading", Proc. Int. Conf. Spoken Language Processing, Philadelphia, PA, 62-65, Oct. 1996.

[4] Z.Y. Wu, S. Zhang, L.H. Cai, H.M. Meng, "Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar", Proc. Int. Conf. Spoken Language Processing, 1802-1805, 2006.

[5] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, "XM2VTSDB: the extended M2VTS database," Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication, 72-77, 1999.

[6] C. Xu, J. Prince, "Snakes, shapes, and gradient vector flow," IEEE Transactions on Image Processing, vol.7, no.3, 1998.

[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, "Active appearance models," Proc. 5th European Conference on Computer Vision, H.Burkhardt and B.Neumann, eds., vol.2, pp.484-498, Springer, Berlin, 1998.

[8] T.F. Cootes, C.J. Taylor, D. Cooper, J. Graham, "Active shape models- their training and application," Computer Vision and Image Understanding, vol.61, pp.38-59, 1995.

[9] A.L. Yuille, P.W. Hallinan, D.S. Cohen, "Feature extraction from faces using deformable templates", Int. J. of Computer Vision, 8(2): 99-111, 1992.

[10] Motion Pictures Expert Group, ISO/IEC 14496-2: 1999/Amd. 1: 2000(E). International Standard, Information Technology – Coding of Audio-Visual Objects. Part 2: Visual; Amendment 1: Visual Extensions.