Robust Lip Tracking by Combining Shape, Color and Motion

Ying-li Tian
Robotics Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213
yltian@cs.cmu.edu
National Laboratory of
Pattern Recognition
Chinese Academy of Sciences,
Beijing, China

Takeo Kanade
Robotics Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213
tk@cs.cmu.edu

Jeffrey F. Cohn
Department of Psychology,
University of Pittsburgh
Pittsburgh, PA 15260
jeffcohn@pitt.edu

Abstract

Accurately tracking facial features requires coping with the large variation in appearance across subjects and the combination of rigid and non-rigid motion. In this paper, we describe our work toward developing a robust method of tracking facial features, in particular, lip contours, by using a multi-state mouth model and combining lip color, shape and motion information. Three lip states are explicitly modeled: open, relatively closed, and tightly closed. The gross shapes of lip contours are modeled by using different lip templates. Given the initial location of the lip template in the first frame, the lip and skin color is modeled by a Gaussian mixture. Several points of a lip are tracked over the image sequence, and the lip contours are obtained by calculating the corresponding lip template parameters. The color and shape information is used to obtain lip states. Our method has been tested on 5000 images from the University of Pittsburgh-Carnegie Mellon University (Pitt-CMU) Facial Expression AU Coded Database, which includes image sequences of children and adults of European, African, and Asian ancestry. The subjects were videotaped while displaying a wide variety of facial expressions with and without head motion. Accurate tracking results were obtained in 99% of the image sequences. Processing speed on a Pentium II 400MHZ PC was approximately 4 frames/second. The multi-state model based method accurately tracked lip motion and was robust to variation in facial appearance among subjects, specularity, mouth state, and head motion.

Keywords: Lip tracking Multiple-state mouth model Lip template

1. Introduction

Robust and accurate analysis of facial features requires coping with the large variation in appearance across subjects and the large appearance variability of a single subject caused by changes in lighting, pose, and facial expressions. Facial analysis has received a great deal of attention in the face detection and recognition literature [1, 2, 6, 14]. Mouth features play a central role for automatic face recognition, facial expression analysis, lip-readings and speech processing. Accurately and robustly tracking lip motion in image sequences is especially difficult because lips are highly deformable, and they vary in shape, color, specularity, and relation to surrounding features across individuals; in addition, they are subject to both non-rigid (expression) and rigid motion (head movement). Although many lip tracking methods have been proposed, there are the limitations to each method when it comes to obtaining robust results.

In this paper, we developed a robust method for tracking lip contours in color image sequence by

combining color, shape and motion. A multi-state mouth model is used to represent the different mouth states: open, relatively closed, and tightly closed. Two parabolic arcs are used as a lip template for modeling the lip shape. Our goal is to robustly track the lip contour in image sequences across individual, specularity and facial expression. The lip template is manually located in the first frame, and the lip color information is modeled as a Gaussian mixture. Then, the key points of the lip template are automatically tracked in the image sequences. The lip states and state transitions are determined by the lip shape and color. We have tested our method on the Pitt-CMU Facial Expression AU Coded Database that includes more than 5000 images of different kinds of people (including Caucasian, Afro-American, Hispanic and Asian) and expressions. Excellent results have been obtained even when there is head motion.

2. Lip Tracking

Each of the lip tracking methods that have been proposed so far has its own strengths and limitations. We believe that a feature extraction system intended to be robust to all the sources of variability (i.e., individual differences in people's appearance, etc.), should use as much knowledge about the scene as possible. Lip tracking methods based on a single cue about the scene are insufficient for robustly and accurately tracking lips. For example, the snake [7] and active contour methods [11] often converge to the wrong result when the lip edges are indistinct or when lip color is very close to face color. Luettin and Thacker [10] proposed a lip tracking method for speechreading using probabilistic models. Their method needs a large set of training data to learn patterns of typical lip deformation. The lip feature point tracking method of Lien [8] is sensitive to the initial feature points position, and the lip feature points have ambiguity along the lip edges. A feature extraction system should use all available information about the scene to handle all the sources of variability in real environments (illumination, individual appearance, etc.).

Many researchers tried to combine more information. Bregler and his colleagues [4] developed an audio-visual speech recognition system that uses Kass's snake approach with shape constraints imposed on possible contour deformations. They found that the outer lip contour was not sufficiently distinctive. This method uses image forces consisting of gray-level gradients, which are known to be inadequate for identifying the outer lip contour [14]. Yuille et al. used the edge, peak and valley information with mouth template for lip tracking, but still experienced some problems during energy minimizing. The weights for each energy term were adjusted by performing preliminary experiments. This process was time-consuming, and the weights were not applicable to the novel subjects. The color-based deformable template method developed by Rao [13] combines shape and color information, but it has difficulty when there is a shadow area near the lip or the lip color is similar to that of the face. Examples of some of these problems are shown in Figure 1. The limitations of these methods can be clearly observed. Figure 1(a) shows the failure of Lien's feature points tracking when the lip contour becomes occluded. The feature points on the lip shift to wrong positions when lips are tightly closed. Figure 1 (b) shows the failure of Rao's color-based deformable method due to shadow and to occlusion. The lower lip contour jumps to the chin because of the shadow.

(a) Lip tracking by the feature point tracking method. Lip contour points shift to the wrong positions.

(b) Lip tracking by the color-based deformable template method. Lower lip contour jumps to the wrong position because of the shadow.

Figure 1. Tracking problems by using different algorithms. The lip contours shift to the wrong positions.

3. Lip Tracking by Combining Color, Shape and Motion

3.1. Multi-state Mouth Model

As shown in Figure 2, we classify the mouth states as open, relatively closed, and tightly closed. We define the lip state as tightly closed if the lips are invisible because of lip suck. For the different states, different lip templates are used to obtain the lip contour (Figure 2 (e), (f), and (g)). For the open mouth, a more complex template could be used that includes inner lip contour and visibility of teeth or tongue. Currently, only the outer lip contour is considered. For the relatively closed mouth, the outer contour of the lips is modeled by two parabolic arcs with six parameters: lip center (xc, yc), lip shape ($h1$, $h2$ and w), and lip rotation (θ). For the tightly closed mouth, the dark mouth line ended at lip corners is used (Figure 2(g)). The state transitions are determined by the lip shape and color.

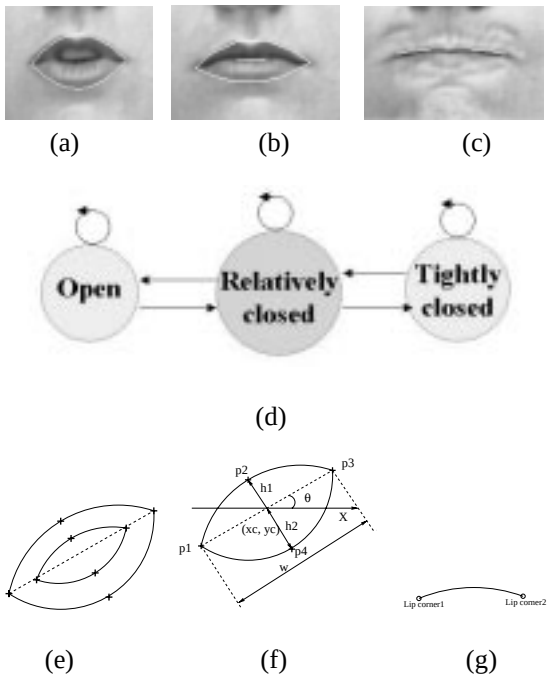


Figure 2. Multi-state mouth model.

3.2. Lip Color Distribution

We model the color distribution inside the closed mouth as a Gaussian mixture. There are three prominent color regions inside the mouth: a dark aperture, pink lips, and bright specularity. The density functions of the mouth Gaussian mixtures are given by:

$$f_{mouth}(x) = \sum_{j=1}^{3} w_j N(x|\mu_j, \mathbf{C}_j), \qquad (1)$$

where

$$N(x|\mu_j, \mathbf{C}_j) = \frac{1}{(2\pi)^{N/2}|\ \mathbf{C}_j\ |^{1/2}}$$
$$exp\{-\frac{1}{2}(x-\mu_j)^T \mathbf{C}_j^{-1}(x-\mu_j)\}. \qquad (2)$$

$\{w_j\}$ are the mixture weights($\sum_j^3 w_j = 1, w_j >= 0$), $\{\mu_j\}$ are the mixture means, and $\{\mathbf{C}_j\}$ are the mixture covariance matrices.

In order to identify the model parameter values, the lip region is manually specified in the first frame image. The Expectation-Maximization (EM) algorithm [5] is used to estimate both the mixture weights and the underlying Gaussian parameters. K-means clustering is used to provide initial estimates of the parameters. Once a model is built, the succeeding frames are tested by a look-up table obtained from the first frame.

3.3. Lip Tracking by Combining Shape and Motion Information

Lip motion: In our system, the lip motion is obtained by a modified version of the Lucas-Kanade tracking algorithm [9]. We assume that intensity values of any given region (feature window size) do not change, but merely shift from one position to another. Consider an intensity feature template $I_t(x)$ over a $n \times n$ region R in the reference image at time t; we wish to find the translation d of this region in the following frame $I_{t+1}(x+d)$ at time $t+1$, by minimizing a cost function E defined as:

$$E = \sum_{x \in R} [I_{t+1}(x+d) - I_t(x)]^2. \qquad (3)$$

and the minimization for finding the translation d can be done in iterations:

$$d_{n+1} = d_n + \left\{ \sum_{x \in R} (\frac{\partial I}{\partial x})^T |_{x+d_n} [I_t(x) - I_{t+1}(x)] \right\} \left[\sum_{x \in R} (\frac{\partial I}{\partial x})(\frac{\partial I}{\partial x})^T |_{x+d_n} \right]^{-1}, \qquad (4)$$

here d_0, the initial estimate, can be taken as zero if only small displacements are involved.

Consecutive frames of an image sequence may contain large feature-point motion such as sudden head movements, brows raised or mouth opening with a the surprised expression; any or all of these may cause missing or lost tracking. In order to track these large motions without losing sub-pixel accuracy, a pyramid method with reduced resolution is used [12]. Each image is decomposed into 5 levels, from level 0 (the original finest resolution image) to level 4 (the coarsest resolution image). In our implementation, a 5x5 Gaussian filter is used to smooth out the noise in order to enhance the computation convergence, and a 13x13 feature region is used for all levels. The rapid and large displacements of up to 100 pixels can be tracked robustly while maintaining sensitivity to sub-pixel facial motion.

Lip template: A lip template is used to obtain the correct lip region. After locating the lip template in the first frame, only four key points of the lip template are tracked (p1, p2, p3 and p4 in Figure 2(f)) in the remainder of the sequence. The lip corners are tracked exactly and the top lip and bottom lip heights are obtained. From the lip corner positions and the lip heights, the lip contour is obtained by calculating the corresponding lip template parameters. The combined method is more robust and accurate for most images, but fails for the tightly closed lip (Figure 5(a)). We solve this problem by combining color information based on a multi-state mouth model.

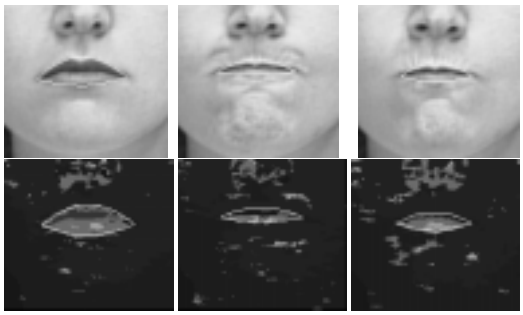
3.4. Lip State Detection by Color and Shape

Lip State Detection by Color and Shape: Each lip state and its color distribution is shown in Figure 3.

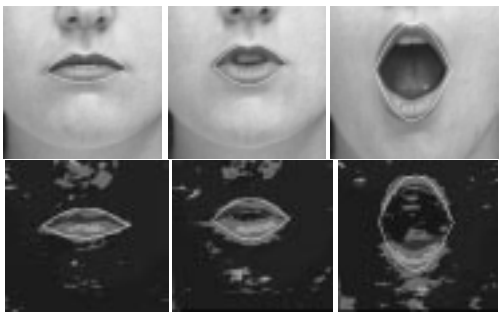
For the open mouth and the tightly closed mouth, there are non-lip pixels inside the lip contours. Assume the lip state in the first frame is neutral closed. From the lip color distribution, we get the lip states by:

$$Lipstate = \begin{cases} Open & \text{if } (h-h_0)/h_0 > 0 \text{ and } \gamma > T_1 \\ Tightly\ closed & \text{if } (h-h_0)/h_0 < 0 \text{ and } \gamma > T_2, \\ Closed & \text{otherwise} \end{cases} \quad (5)$$

where h_0 and h are the sum of top lip and bottom lip heights in the first frame and current frame; $\gamma = n_{nonlip}/n$, n_{nonlip} is non-lip pixel number inside lip contour and n is all the pixel numbers inside the lip contour; $T_1 = 0.35$ and $T_2 = 0.25$ are thresholds in our application.



(a) The original image and color distribution of tightly closed mouth



(b) The original image and color distribution of open mouth

Figure 3. Obtaining Lip states from lip color information. For the open mouth and the tightly closed mouth, there are non-lip pixels inside the lip contours.

Lip Contours for Tightly Closed Lips: For the tightly closed mouth, we trace the mouth line by locating the darkest pixels along the perpendicular lines with a distance r between two lip corners (Figure 4).

Based on the multi-state mouth model, the color, shape, and motion information are combined in our lip tracking method. This method is very robust to tracking lips for each state. Figure 5(a) shows the lip tracking results by using shape and motion information, but not using the multi-state mouth model. The lip contours are not correct for tightly closed lip. Figure 5(b) shows the correct lip tracking results for tightly closed lip by adding the multi-state mouth model.

(a) Approximate contour (b) Refined contour

Figure 4. Obtaining tightly closed lip line.

(a) Lip tracking by combining shape and motion without using the multi-state mouth model. The lip contours for tightly closed lips are not correct.

(b) Lip tracking by combining shape, color and motion based on multi-state mouth model. The lip contours for tightly closed lips are correct.

Figure 5. Tracking results by OUR METHOD. Correct lip contours are obtained for tightly closed.

4. Experiment Results

We tested our method on 500 image sequences (approximately 5000 images) from the Pitt-CMU Facial Expression AU Coded Database. Subjects ranged in age from 3 to 30 years old and included males and females of European, African, and Asian ancestry. They were videotaped in an indoor environment with uniform lighting. During recording, the camera was positioned in front of the subjects and provided for a full-face view. Images were digitized into 640x480 pixel arrays with 24-bit color resolution. A large variety of lip motions was represented [3].

Figures 6, 7,and 8 show representative results of lip

contour tracking. Figure 6 shows tracking results for the same subject showing different expressions. From

Figure 6. Tracking results by our method for different expressions(happy, surprise, sad, fear, disgust and anger).

the happy, surprised and anger expression sequences, we see that lip states transit each other and correct lip contours are obtained for each lip state. The results for the dark-skinned subjects are shown in Figure 7. Our method works well even for big lip movement, for example, the expression of surprise. Figure 8 demonstrates the robustness of our method for tracking lip contours when subjects have both non-rigid motion and rigid motion. The lip contours are tracked correctly with different kinds of head motion. The first and second rows demonstrate the good lip tracking results for different expressions with horizontal and vertical head rotation. In the last row, the excellent tracking results for the infant are given, including both big lip deformation, big head motion and background motion. Our system can track 4 images/sec on a Pentium II 400MHZ PC and works robustly for image of men, women, and children, and for people of varying skin color and appearance.

Figure 7. Tracking results by our method for darkskin subjects with different expressions.

5. Conclusion and Discussion

We have described a robust method for tracking unadorned lip movements in color image sequences of various subjects and various expressions by combining color, shape and motion information. A multi-state mouth model was introduced. Three lip states were considered: open, relatively closed, and tightly closed. The lip state transitions were determined by the lip shape and color. Given the initial location of the lip template in the first frame, the lip color information was obtained by a Gaussian mixture. Then, the lip key points were tracked via the tracking method developed by Lucas-Kanade[9] in the image sequence and the lip contours were obtained by calculating the corresponding lip template parameters. The color and shape information is used to obtain lip states. Compared with other approaches, our method requires no training data, and works well for different individuals, for mouths with specularity, and for different mouth states. Our method is able to track lips with vertical and horizontal head rotation.

A limitation of our method is that we assume that

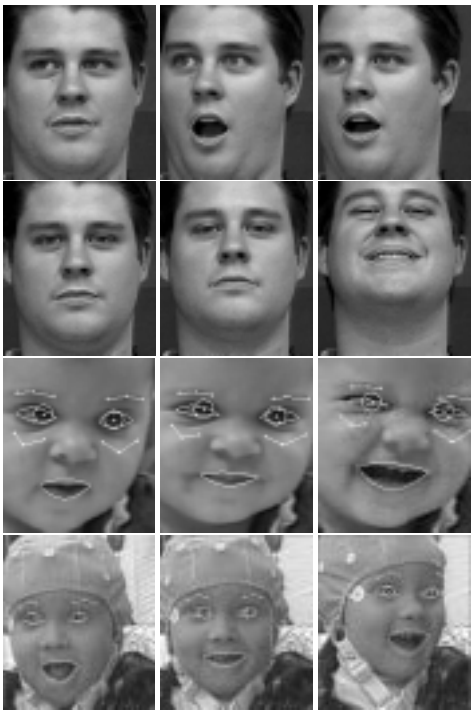


Figure 8. Tracking results by our method for different subjects.

the lip template is symmetrical about the perpendicular bisector to the line connecting the lip corners. For non-symmetrical expressions and complex lip shapes, there are some errors between the tracking lip contour and the actual lip shape (Figure 9). A more complex lip template will be necessary to get more accurate lip contours for non-symmetrical expression analysis in our future work.

(a) (b)

Figure 9. Non-symmetrical expression and complex Lip shapes. "+" indicate the real lip contour

Acknowledgements

The authors would like to thank Zara Ambadar for testing the method on the database. We also thank Simon Baker for his comments and suggestions on earlier versions of this paper. This work is supported by NIMH grant R01 MH51435.

References

[1] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct. 1993.

[2] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[3] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36:35–43, 1999.

[4] J. D. Cowan, G. Tesauro, and J. Alspector(eds). *Surface Learning with Applications to Lipreading*. Advances in Neural Information Processing Systems 6, Morgan Kaufmann Publishers, San Francisco, 1994.

[5] A. Dempster, M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc.,*, (B(39)):1–38, 1977.

[6] L. Huang and C. W. Chen. Human facial feature extraction for face interpretation and recognition. *Pattern Recognition*, 25(12):1435–1444, 1992.

[7] M. Kass, A. Witkin, and D. Terzopoulus. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[8] J.-J. J. Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*. PhD Thesis, University of Pittsburgh, 1998.

[9] B. Lucas and T. Kanade. An interative image registration technique with an application in stereo vision. In *The 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[10] J. Luettin and N. A. Tracker. Speechreading using probabilistic models. *Computer vision and Image Understanding*, 65(2):163–178, Feb. 1997.

[11] J. Luettin, N. A. Tracker, and S. W. Beet. *Active Shape Models for Visual Speech Feature Extraction*. Electronic Systems Group Report No. 95/44, University of Sheffield, UK, 1995.

[12] C. Poelman. The paraperspective and projective factorization method for recovering shape and motion. *Technical Report CMU-CS-95-173, Carnegie Mellon University*, 1995.

[13] R. R. Rao. *Audio-Visal Interaction in Multimedia*. PHD Thesis, Electrical Engineering, Georgia Institute of Technology, 1998.

[14] A. Yuille, P. Haallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision,*, 8(2):99–111, 1992.