Facial Feature Tracking using a Multi-State Hierarchical Shape Model under Varying Face Pose and Facial Expression

Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA

Abstract

This paper presents a multi-state hierarchical approach for facial feature tracking. A hierarchical formulation of statistical shape models is proposed to characterize both global shape constraints of human faces and local structural details of facial components. Gabor wavelets and gray level profiles are integrated for effective and efficient representation of feature points. Furthermore, multi-state local shape models are presented to deal with shape variations of facial components. Meanwhile, face pose estimation helps improve shape constraints for the feature search. Both facial component states and feature point positions are dynamically estimated using a multi-modal tracking approach. Experimental results demonstrate that the proposed method accurately and robustly tracks facial features under different facial expressions and pose variations.

1. Introduction

Facial feature localization and tracking is important in application areas such as man-machine interaction, human identification, and expression recognition. However, accurate and efficient tracking of facial feature points could be a tough issue due to the potential variability such as facial expression change and pose variations in image sequences.

Extensive work has been focused on the shape representation of deformable objects, such as Active Contour Model (Snake)[9], Elastic Bunch Graph Matching (EBGM)[16], Active Shape Model (ASM)[6], and Active Appearance Model (AAM)[3]. Gabor wavelets have also been employed as feature representation to improve the robustness and accuracy of feature point search [11]. Moreover, to deal with nonlinear shape space caused by pose variation, [5][4][2][8] use a collection of 2D local linear models, [10][1][17] explicitly employ a 3D face model and [12] is utilizing a nonlinear kernel PCA model. To handle appearance variations due to facial expression changes, [13] proposes a multi-state facial component model combining the color, shape, and motion information.

This paper presents a multi-state hierarchical shape model for facial feature tracking in near frontal-view and half profile-view image sequences. Based on the active shape model, a two-level hierarchy in feature points is proposed to characterize global shape of human faces, and local structural details of facial components. Gabor wavelet jets and gray level profiles are combined to represent the feature points in an effective and efficient way. Multi-state local shape models are further presented to deal with shape variations of facial components. Moreover, the use of three-dimensional face pose estimation improves shape constraints for the feature search. Both states of facial components and positions of feature points are dynamically estimated by a multi-modal tracking approach.

The rest of the paper is arranged as follows: Section 2 proposes the multi-state hierarchical shape model, Section 3 presents the multi-state facial tracking algorithm, Section 4 exhibits the experimental results, then our technique is concluded in Section 5.

2. Multi-state Hierarchical facial shape model

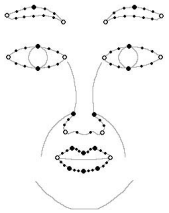


Figure 1. Feature points in the facial model: fiducial points marked by circles (global) and big black dots (local), and contour points (local) marked by small black dots.

To represent deformable objects, statistical shape models assume that the structure for a specific class of objects can be represented by a set of feature points. Figure 1 shows the layout of feature points in our facial model including fiducial points controlling the shape deformation of human faces (e.g. eye corners and mouth corners) and contour points interpolated between the fiducial points along the boundary.

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

In the ASM, all the feature point positions are updated (or projected) simultaneously, which indicates that the interactions within feature points are simply parallel. However, this may not be adequate to describe the sophisticated structure of human faces. For example, given the corner points of an eye, whether the eye is open or close will not affect the localization of mouth or nose. Generally, facial feature points can be organized into two categories: global feature points characterizing the global shape constraints for the entire face, and local feature points capturing the local structural details. Based on this two-level hierarchy, a hierarchical formulation of statistical shape model is presented to characterize possible shape variations of human faces.

The face shape vector $\mathbf{s}$ could be expressed as $(\mathbf{s}_g, \mathbf{s}_l)^T$, where $\mathbf{s}_g$ and $\mathbf{s}_l$ denote global and local feature points respectively as shown in Figure 1. The facial model is partitioned into four components: eyebrows, eyes, nose, and mouth. The two eyes (or eyebrows) are considered as one facial component because of their symmetry.

For the global face shape, a point distribution model can be learned from the training data.

$$\mathbf{s}_g = \bar{\mathbf{s}}_g + \mathbf{P}_g \mathbf{b}_g \tag{1}$$

where $\bar{\mathbf{s}}_g$ is the mean global shape, $\mathbf{P}_g$ is a set of principal orthogonal modes of global shape variation, and $\mathbf{b}_g$ is a vector of global shape parameters.

Since it is difficult for a single-state statistical shape model to handle nonlinear face deformations due to expression changes, multi-state local shape models are further presented to deal with shape variations of facial components. In our facial model, there are three states (open, closed, and tightly closed) for the mouth, two states (open and closed) for the eyes, and one state for the other two components (eyebrows and nose). Since the global feature points are relatively less influenced by local structural variations, it is assumed that the state switching of facial components only involves local shape models. For the ith facial component under the jth state, the shape vector becomes $\mathbf{s}_{g_i,l_{i,j}} = (\mathbf{s}_{g_i}, \mathbf{s}_{l_i})^T$. The local shape model is expressed by the multi-state formulation:

$$\mathbf{s}_{g_i,l_{i,j}} = \bar{\mathbf{s}}_{g_i,l_{i,j}} + \mathbf{P}_{g_i,l_{i,j}} \mathbf{b}_{g_i,l_{i,j}} \tag{2}$$

where $\bar{\mathbf{s}}_{g_i,l_{i,j}}$, $\mathbf{P}_{g_i,l_{i,j}}$, and $\mathbf{b}_{g_i,l_{i,j}}$ are the corresponding mean shape, principal shape variation, and shape parameters. Thus the multi-state hierarchical shape model consists of a global shape model and a set of multi-state local shape models.

3. Multi-state facial feature tracking

For each facial component, both the component state and positions are tracked by a dynamic multi-modal approach.

3.1. Multi-modal component tracking

In this work, a switching hypothesized measurements (SHM) model [15] is applied to enable switching the component state $\mathbf{s}_t$ and to estimate the hidden state $\mathbf{z}_t$ (i.e. feature point positions) of each facial component at time instant t. The hidden state transition can be modeled as

$$\mathbf{z}_{t+1} = \mathbf{F}\mathbf{z}_t + \mathbf{n} \tag{3}$$

where $\mathbf{F}$ is the state transition matrix, and $\mathbf{n}$ represents the system perturbation. The switching state transition is modeled by a first order Markov process to encourage the temporal continuity of the component state. Since the component state is unknown, feature point positions are searched once under each possible component state. Under the assumption of the jth state of the component, a hypothesized measurement $\mathbf{o}_{t,j}$ represents the feature point positions of the facial component obtained from the feature search procedure at time t. Given s_t, the corresponding hypothesized measurement $\mathbf{o}_{t,s_t}$ could be considered as a proper measurement centering on the true feature positions, while every other $\mathbf{o}_{t,j}$ for $j \neq \mathbf{s}_t$ is an improper measurement generated under a wrong assumption. The improper measurement should be weakly influenced by true feature positions and have a large variance. To simplify the computation, the measurement model is formulated as

$$\mathbf{o}_{t,j} = \begin{cases} \mathbf{H}\mathbf{z}_t + \mathbf{v}_{t,j} & if\ j = s_t \\ \mathbf{w} & otherwise \end{cases} \tag{4}$$

where $\mathbf{H}$ is the measurement matrix, $\mathbf{v}_{t,j}$ represents the measurement uncertainty assuming as a zero mean Gaussian, and $\mathbf{w}$ is a uniformly distributed noise. The state transition matrix $\mathbf{F}$, system noise $\mathbf{n}$, and measurement matrix $\mathbf{H}$ are defined in the same way as in a Kalman filter [7].

Given the state transition model, measurement model, and measurement data, the hidden state $\mathbf{z}_t$ as well as the switching state s_t can be recursively estimated by the SHM filtering algorithm [15]. Since the measurement under the true hypothesis of the switching state usually shows more regularity and has smaller variance compared with the other hypothesized measurements, the true information (the facial component state and feature point positions) could be enhanced through the propagation in the SHM filter. Moreover, for a facial component with only one state, the multi-modal SHM filter degenerates into a unimodal Kalman filter.

3.2. Facial tracking algorithm

For each frame, the face region is normalized to a 64×64 image using the knowledge of eye centers obtained by an eye detector [14]. Given the multi-state hierarchical shape model, the facial feature tracking algorithm performs an iterative process at time t:

- Project the global mean shape $\bar{\mathbf{s}}_g$, and the local mean shape $\bar{\mathbf{s}}_{g_i,l_{i,j}}$ using the estimated face pose from the previous frame. The modified mean shapes are more suitable for the current pose and provide better shape constraints for the feature search.

- Localize the global feature points individually, and update the global shape parameters to match $\mathbf{s}_g$ with the constraints on $\mathbf{b}_g$.
- Generate $\mathbf{s}_g$ as Eq. (1), and then return to previous step until convergence.

  Enumerate all the possible states of the ith facial components. Under the assumption of the jth state:
  - Localize the local feature points individually ,and update the local shape parameters to match $\mathbf{s}_{g_i,l_{i,j}}$ with the constraints on $\mathbf{b}_{g_i,l_{i,j}}$.
  - Generate the shape vector $\mathbf{s}_{g_i,l_{i,j}}$ as Eq. (2) and then return to previous step until convergence.
- Take the feature search results $(\mathbf{s}_{g_i}, \mathbf{s}_{l_{i,j}})^T$ for the ith facial component under different sate assumptions, as the set of hypothesized measurements $\mathbf{o}_{t,j}$. Estimate the state of the ith facial component and the positions of its feature points at time t through the SHM filter.
- Estimate the $3D$ face pose by the tracked global facial feature points $\mathbf{s}_g$.

3.3. Hybrid facial feature representation

Two different means of feature representation, multi-scale and multi-orientation Gabor wavelet jets [16] and gray level profiles normal to the object boundary [6], are utilized in this work to model the local information of fiducial points and contour points respectively. Gabor wavelet based feature representation provides rich information of local appearances and lead to accurate feature point localization, but with relatively high computation complexity. The profile based representation is computationally efficient, but is not sufficient to identify all the facial feature points. Compared to wavelet based representation for all the feature points, the hybrid representation achieves similar feature search accuracy and enhances the computation speed by 60% in our experiments.

4. Experimental results

The global shape model, multi-state local shape models, Gabor wavelet jets of fiducial points, and gray level profiles of contour points are trained using 500 images from two hundred persons with different races, ages, and expressions. Both feature point positions and facial component states are manually labeled in each training image. The principal orthogonal modes in the shape models stand for 95% of the shape variation. The testing sequences consist of ten sequences, each of which consists 100 frames. The test sequences contain six subjects, which are not included in training data. Our C++ program can process about seven frames per second on a Pentium 4 2.8GHz PC.

Figure 2-5 exhibit the results of facial feature tracking under pose variations and face deformations. To make the results clear, only the fiducial points are shown, since they are more representative than the contour points. Figure 2 shows the results by the proposed method with and without modified mean shapes. Compared to the results in Figure 2(b,d), the feature points are more robustly tracked in Figure 2(a,c) under large pose variations, which demonstrates that the projection of mean shapes helps improve shape constraints. Figure 3 shows the results by the proposed method with and without the multi-state local shape models. It can be seen from Figure 3 that the multi-state models substantially improve the robustness of facial feature tracking, especially when eyes and mouth open or close. That demonstrates the state switching in local shape models helps deal with nonlinear shape deformations of facial components.

(a) (b) (c) (d)

Figure 2. Feature tracking results: (a,c) proposed method, (b,d) without using modified mean shapes.

(a) (b) (c) (d)

Figure 3. Feature tracking results: (a,c) proposed method, (b,d) without using the multi-state local shape models

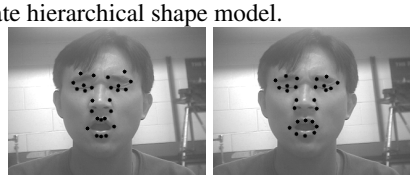
(a) (b) (c) (d)

Figure 4. Feature tracking results: (a,c) proposed method, (b,d) without using the hierarchical shape model

Figure 4 shows the results by the proposed method with and without the hierarchical shape model. Compared to the results in Figure 4(b,d), the feature points are more accurately tracked in Figure 4(a,c) using the hierarchical shape model, since the two-level hierarchy in the facial shape model provides a relatively sophisticated structure to describe the interactions among feature points. Figure 5 compares the feature tracking results of the proposed method with the results of the elastic bunch graph matching approach respectively. The mean shapes are projected to the estimated face pose for both the methods. It can be seen that overall the proposed multi-state hierarchical method outperforms EBGM under pose and expression variations.

The results of facial feature tracking are evaluated quantitatively besides visual comparison. Fiducial and contour feature points are manually marked in 1,000 frames from the ten sequences for comparison. For each feature point,

the displacement (in pixels) between the estimated position and the corresponding labeled position is computed as the feature tracking error. Figure 6 shows error distribution of the feature tracking results by different methods. Compared to the feature tracking result using single state local model and using multi-state local models without hierarchical models, the use of the multi-state hierarchical shape model averagely reduce the feature tracking error by 24% and 13% respectively. Moreover, besides the comparison of the average pixel displacement, Figure 7 illustrates the evolution of the error over time for one image sequence, which demonstrates the proposed method substantially improves the robustness of facial feature tracking using the multi-state hierarchical shape model.



(a) (b)

Figure 5. Feature tracking results: (a) proposed method. (b) EBGM approach

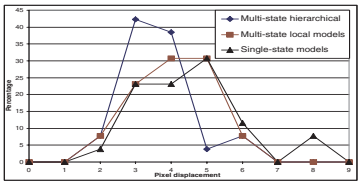


Figure 6. Error distribution of feature tracking. Diamonds: proposed method. Triangles: without the multi-state local shape models. Squares: without the hierarchical shape model.

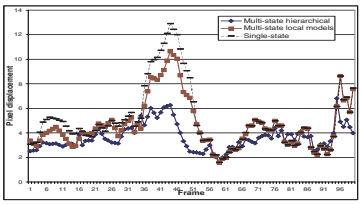


Figure 7. Error evolution of feature tracking for an image sequence. Diamonds: proposed method. Bars: without the multi-state local shape models. Squares: without the hierarchical shape model.

5. Conclusion

In this paper, there are four main contributions for the proposed facial tracking approach. First, a multi-state hierarchical shape model is presented to characterize the global shape constraints and local structural details of human faces. Second, Gabor wavelet jets and gray level profiles are integrated for effective and efficient feature representation. Third, mean shapes are modified through robust face pose estimation to improve shape constraints for the feature search. Fourth, feature point positions are dynamically estimated with multi-state local shape models using a multi-modal tracking approach. Experimental results show that the proposed method significantly improves the accuracy and robustness of facial feature tracking under pose variations and face deformations. Future work will be focused on facial feature detecting and tracking in sequences with profile-view images.

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *SIGGRAPH99*, pages 187–194, 1999.

[2] C. M. Christoudias and T. Darrell. On modelling nonlinear shape-and-texture appearance manifolds. *Proc. of CVPR05*, 2:1067–1074, 2005.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance model. *Proc. of 5th ECCV*, 2:484–498, Jun 1998.

[4] T. F. Cootes and C. J. Taylor. A mixture model for representing shape variation. *Image and Vision Computing*, 17(8):567–574, 1999.

[5] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and Vision Computing*, 20:657–664, 2002.

[6] T. F.Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models —their training and application. *CVIU*, 61(1):38–59, Jan 1995.

[7] H. Gu, Q. Ji, and Z. Zhu. Active facial tracking for fatigue detection. *Proc. IEEE Workshop on Applications of Computer Vision*, pages 137–142, 2002.

[8] T. Heap and D. Hogg. Wormholes in shape space: Tracking through discontinuous changes in shape. *Proc. of ICCV98*, pages 344–349, 1998.

[9] M. Kass, A. WItkin, and D. Terzopoulos. Snakes: Active contour models. *Proc. of ICCV87*, pages 259–268, 1987.

[10] Y. Li, S. Gong, and H. Liddell. Modelling faces dynamically across views and over time. *Proc. of ICCV01*, pages 554–559, Feb 2001.

[11] S. J. McKenna, S. Gong, R. P. Wurtz, J. Tanner, and D. Banin. Tracking facial feature points with gabor wavelets and shape models. *Proc. Int'l Conf. Audio and Video based Biometric Person Authentication*, pages 35–42, 1997.

[12] S. Romdhani, S. Gong, and A. Psarrou. Multi-view nonlinear active shape model using kernel pca. *Proc. of BMVC*, pages 483–492, Sep 1999.

[13] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on PAMI*, 23(2):97–115, Feb 2001.

[14] P. Wang and Q. Ji. Learning discriminant features for multi-view face and eye detection. *Proc. of CVPR05*, 1:373–379, June 2005.

[15] Y. Wang, T. Tan, and K.-F. Loe. Joint region tracking with switching hypothesized measurements. *Proc. of ICCV03*, 1:75–82, 2003.

[16] L. Wiskott, J. Fellous, N. Krger, and C. V. der Maksburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on PAMI*, 19(7):775–779, Jul 1997.

[17] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. *Proc. of CVPR04*, 2:535–542, 2004.