Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 40 (2007) 3195–3208

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/pr

Robust facial feature tracking under varying face pose and facial expression

Yan Tong[a], Yang Wang[b], Zhiwei Zhu[c], Qiang Ji[a,*]

[a]*Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA*
[b]*National ICT Australia, Eveleigh, NSW 1430, Australia*
[c]*Sarnoff Corporation, Princeton, NJ 08543-5300, USA*

Received 22 October 2006; received in revised form 23 February 2007; accepted 28 February 2007

Abstract

This paper presents a hierarchical multi-state pose-dependent approach for facial feature detection and tracking under varying facial expression and face pose. For effective and efficient representation of feature points, a hybrid representation that integrates Gabor wavelets and gray-level profiles is proposed. To model the spatial relations among feature points, a hierarchical statistical face shape model is proposed to characterize both the global shape of human face and the local structural details of each facial component. Furthermore, multi-state local shape models are introduced to deal with shape variations of some facial components under different facial expressions. During detection and tracking, both facial component states and feature point positions, constrained by the hierarchical face shape model, are dynamically estimated using a switching hypothesized measurements (SHM) model. Experimental results demonstrate that the proposed method accurately and robustly tracks facial features in real time under different facial expressions and face poses.

Keywords: Facial feature detection and tracking; Active shape model; Face pose estimation

1. Introduction

Face plays an essential role for human communication. It is the main source of information to discriminate and identify people, to interpret what has been said by lipreading, and to understand one's emotion and intention based on the emotional facial expressions. The facial feature points are the prominent landmarks surrounding facial components: eyebrows, eyes, nose, and mouth. They encode critical information about facial expression and head movement. Therefore, facial feature motion can be defined as a combination of rigid head motion and nonrigid facial deformation. Accurate localization and tracking facial features are important in applications such as vision-based human–machine interaction, face-based human identification, animation, entertainment, etc. Generally, the facial feature tracking technologies could be classified into two categories: model-free and model-based tracking algorithms. The model-free tracking algorithms [1–7] are general purpose point trackers without the prior knowledge of the object. Each facial feature point is usually tracked by performing a local search for the best matching position, around which the appearance is most similar to the one in the initial frame. However, the model-free methods are susceptible to the inevitable tracking errors due to the aperture problems, noise, and occlusion. Model-based methods, on the other hand, focus on explicit modeling the shape of the objects. Recently, extensive work has been focused on the shape representation of deformable objects such as active contour models (Snakes) [8], deformable template method [9], active shape model (ASM) [10], active appearance model (AAM) [11], direct appearance model (DAM) [12], elastic bunch graph matching (EBGM) [13], morphable models [14], and active blobs [15]. Although the model-based methods utilize much knowledge on face to realize an effective tracking, these models are limited to some common assumptions, e.g. a nearly frontal view face and moderate facial expression changes, and tend to fail under large pose variations or facial deformations in real-world applications.

Given these challenges, accurate and efficient tracking of facial feature points under varying facial expression and face pose remains challenging. These challenges arise from the potential variability such as nonrigid face shape deformations caused by

[*] Corresponding author. Tel.: +1 518 2766440.
E-mail address: qji@ecse.rpi.edu (Q. Ji).

0031-3203/$30.00 © 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.02.021

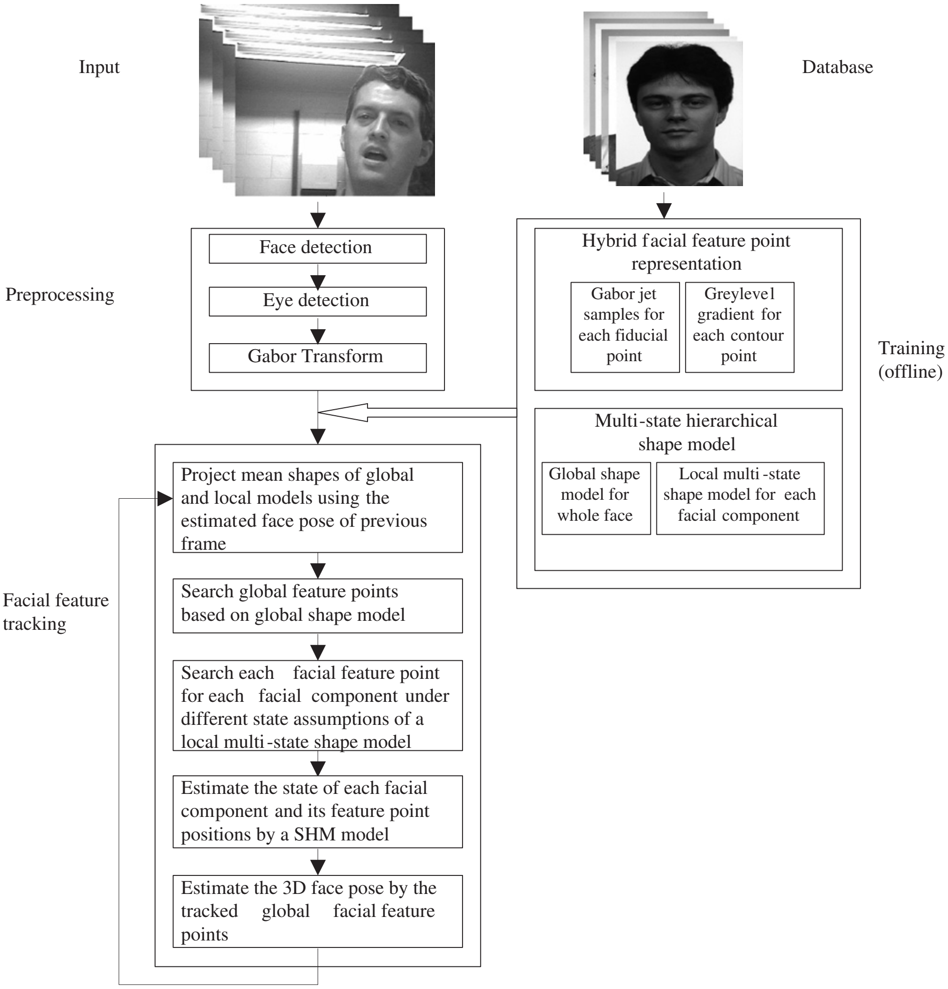
Input
Database
Preprocessing
Face detection
Eye detection
Gabor Transform
Hybrid facial feature point representation
Gabor jet samples for each fiducial point
Greylevel gradient for each contour point
Training (offline)
Multi-state hierarchical shape model
Global shape model for whole face
Local multi-state shape model for each facial component
Facial feature tracking
Project mean shapes of global and local models using the estimated face pose of previous frame
Search global feature points based on global shape model
Search each facial feature point for each facial component under different state assumptions of a local multi-state shape model
Estimate the state of each facial component and its feature point positions by a SHM model
Estimate the 3D face pose by the tracked global facial feature points

Fig. 1. The flowchart of the automatic facial feature tracking system based on the multi-state hierarchical shape model.

facial expression change, the nonlinear face transformation resulting from pose variations, and illumination changes in real-world conditions. Tracking mouth and eye motion in image sequences is especially difficult, since these facial components are highly deformable, varying in both shape and color, and subject to occlusion.

In this paper, a multi-state pose-dependent hierarchical shape model is presented for facial feature tracking under varying face pose and facial expression. The flowchart in Fig. 1 summarizes our method. Based on the ASM, a two-level hierarchical face shape model is proposed to simultaneously characterize the global shape of a human face and the local structural details of each facial component. Multi-state local shape models are further introduced to deal with shape variations of facial components. To compensate face shape deformation due to face pose change, a robust 3D pose estimation technique is introduced, and the hierarchical face shape model is corrected based on the estimated face pose to improve the effectiveness of the shape constraints under different poses. Gabor wavelet jets and gray-level profiles are combined to represent the feature points in an effective and efficient way. 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 provides a detailed review on the related work of model-based facial feature tracking approaches. Section 3 presents our proposed facial feature tracking algorithm including the hierarchical multi-state pose-dependent face shape model, the hybrid feature representation, and the proposed multi-modal facial feature tracking algorithm. Section 4 discusses the experimental results. The paper concludes in Section 5, with a summary and discussion for future research.

2. Related work

2.1. Facial feature tracking in nearly frontal view

Extensive recent work in facial component detection and tracking has utilized the shape representation of deformable objects, where the facial component shape is represented by a set of facial feature points.

Wiskott et al. [13] present the EBGM method to locate facial features using object adopted graphs. The local information
]

of feature points is represented by Gabor wavelets, and the geometry of human face is encoded by edges in the graph. The facial features are extracted by maximizing the similarity between the novel image and model graphs.

Recently, statistical models have been widely employed in facial analysis. The ASM [10] proposed by Cootes et al., is a popular statistical approach to represent deformable objects, where shapes are represented by a set of feature points. Feature points are searched by gray-level profiles, and principal component analysis (PCA) is applied to analyze the modes of shape variation so that the object shape can only deform in specific ways that are found in the training data. Robust parameter estimation and Gabor wavelets have also been employed in ASM to improve the robustness and accuracy of feature point search [16,17]. Instead of using gray-level profiles to represent feature points, multi-scale and multi-orientation Gabor wavelet coefficients are utilized to characterize the local appearance around feature points. The AAM [11] and DAM [12] are subsequently proposed to combine constraints of both shape variation and texture variation.

Unfortunately, the current statistical model-based facial feature detection and tracking methods are limited to a narrow scope due to the global linear assumptions with PCAs. Research has shown that the assumption of nearly frontal view is effective and convenient in facial feature tracking, while it tends to fail under large pose variations or significant facial expressions in real-world applications.

2.2. Facial feature tracking under varying pose and expressions

Recent research has been dedicated to model the nonlinearity of facial deformations caused by pose variations or facial expressions. Generally, these approaches could be grouped into three categories: a group of approaches utilizes a collection of local linear models to deal with the global nonlinearity [18–21], an alternate class of technologies employs 3D facial models derived from the image sequences [22–24], and another group of approaches [25,26] models the nonlinearity explicitly.

(1) *Multi-modal approaches for facial feature tracking*: The multi-modal approach assumes that each shape/appearance corresponding to one specific pose (or expression) could be approximated linearly by a single shape/appearance model, such that a set of 2D linear shape/appearance models could be combined to model the nonlinear variations. Cootes et al. proposed a weighted mixture of Gaussian models [20] to represent the complex shape variation. An EM algorithm is used to estimate the model parameters from a set of training data. After the mixture Gaussian model is obtained, PCA is applied to each Gaussian component for dimensional reduction. Similar to the conventional ASM, feature search is performed once for each mixture component. By projecting back into the original shape space, the probability that the searched shape is generated by the model is computed, and the component that yields the highest probability is selected. Christoudias et al. [21] extend the mixture Gaussian model to represent both the object appearance and shape manifold in image space. The mixture model restricts its search to valid shapes and appearance, and therefore avoids erroneous matches. However, the main problem for both methods is that performing mixture of Gaussian fitting in original space is time consuming and requires a lot of training data due to high dimensionality with each component, if each component is described by a full covariance matrix. In addition, the best facial features are detected by enumerating each mixture component, which is again time consuming and makes real-time implementation of such methods infeasible.

Assuming that the model parameters are related to the face pose, Cootes et al. develop a view-based AAM [18] to represent the face from a wide range of face poses. The view-based AAM consists of five 2D shape models, each of which represents the shape deformation from a specific view point. Each 2D model is trained using a different set of feature points from a set of training images taken within a narrow range of head pose for each view. The relationship between face pose angle and the model parameters can be learned from images taken from different views simultaneously. Initially, the best match is achieved by comparing the searching results against the models from all view points. The head pose is then estimated from the model parameters, and the facial features are tracked given the head pose. However, only one pose parameter (pan, tilt or swing) is considered at one time, so that the feature points are assumed to move along circles in 3D space and ellipses in 2D. In real-world condition, however, the head rotation often involves a combination of the three angles. In addition, enumerating each view to find the best view is also time consuming.

Similarly, Yan et al. [27] extend the DAM into multi-view application by combining several DAMs, each of which is trained from a range of face poses. During the facial feature tracking process, the models corresponding to the previous view and the neighboring two views are attempted, and the best matching is chosen as the one with the minimum texture residual error.

Grauman et al. [28] propose a nearest neighbor (NN) method to represent the human body shape across poses. Christoudias et al. [21] extend the NN method to model the manifolds of facial features. Starting from an initial guess, the NN search is performed by minimizing the distance with all prototype examples in pixel space and retaining the k-nearest neighbors, then a new example could be generated by using a convex combination of the neighborhood's shape and texture. This method has several advantages: it does not need assumption about the global structure of the manifold, and it could be more naturally extended to shape features having multiple dimensionality. However, it needs many representative prototype examples.

The hierarchical point distribution model (PDM) is proposed by Heap et al. [19] to deal with the highly nonlinear shape deformation, which will result in discontinuous shape parameter space generated by PCA. The hierarchical PDM consists of two levels of PCAs: in the first level, the training data is projected into shape parameter space, and the shape parameter space is divided into several groups, each of which corresponds to a distinct face pose and facial expression combination, by clustering; in the second level, each cluster is projected onto a local PCA space, respectively, to give a set of overlapped local hyper ellipsoids. A highly nonlinear shape parameters space is generated by the union of the piece-wise local PCAs. Different from the mixture Gaussian approaches, the hierarchical PDM does not have a probabilistic framework. In addition, choosing the optimal number of local clusters is difficult, and substantial data is required for constructing the model. Although hierarchical formulations are employed in Ref. [19], all the feature points are at the same level, and their positions are updated simultaneously.

To handle appearance variations caused by the facial expression changes, Tian et al. [29] propose a multi-state facial component model combining the color, shape, and motion information. Different facial component models are used for the lip, eyes, brows, respectively. Moreover, for lip and eyes, each component model has multiple states with distinguished shapes. The state of component model is selected by tracking a few control points. The accuracy of their method, however, critically relies on how accurately and reliably the control points are tracked in current frame. Hence, an automatic state switching strategy is desirable. In addition, the facial feature points are manually initialized in the first frame.

The multi-modal approaches have the advantage of handling very large face pose by using a few linear statistical models with different topologies or even different dimensionality for specific view points. These approaches could also be generalized to deal with the nonlinear deformation due to facial expressions. The main weakness with current multi-modal approaches is that they could handle facial deformation either from significant facial expression change or from face pose variation, but not both. In contrast, our proposed hierarchical model explicitly accounts for the deformations of the facial components under different states and poses, therefore is able to track the facial features under large variation of facial expression and face pose.

(2) *Modeling facial features by using a* 3D *face model*: The previous approaches on 3D deformable models [30–32] utilize a 3D face mesh to model the global rigid motion and the local nonrigid facial motion, respectively, by a two-step procedure. The 3D local models, however, do not explicitly model each facial component. Furthermore, the complexity to obtain a dense point-to-point correspondence between vertices of the face and the face model is considerable. In addition, the local models used in the 3D deformable model are not sufficient to handle the high nonlinearity due to significant facial expression changes.

To reduce the computation complexity, Li et al. [22] propose a multi-view dynamic sparse face model to model 3D face shape from video sequence, instead of a dense model. The model consists of a sparse 3D facial feature PDM, a shape-and-pose-free texture model, and an affine geometrical model. The 3D face shape model is learned from a set of facial feature points from 2D images with labeled face poses. The 3D coordinates of each facial feature point are estimated using an orthographic projection model approximation. The texture model is extracted from a set of shape-and-pose-free face images, which is obtained by warping the face images with pose changes onto the mean shape at frontal view. The affine geometrical model is used to control the rotation, scale, and translation of faces. The fitting process is performed by randomly sampling the shape parameters around the initial values. The best matching set of parameters is obtained by evaluating a loss function for all possible combinations. The time complexity of such a method is high, and it could not handle facial deformation due to facial expression change.

Xiao et al. [24] propose a 2D + 3D AAM, in which they extend the 2D AAM to model 3D shape variation with additional shape parameters and use a nonrigid structure-from-motion algorithm [33] to construct the corresponding 3D shape modes of the 2D AAM. Compared with the 3D morphable model with a dense 3D shape model and multi-view dynamic face model with a sparse 3D shape model, the 2D+3D AAM only has 2D shape model. The 2D shape model has the capability to represent the same 3D shape variations as 3D shape model, while needing more shape parameters. However, the 2D shape model can generate the shape modes, which are impossible for 3D model. The 3D pose obtained by a structure-from-motion method is used to constrain the 2D AAM to only generate the valid shape modes corresponding to possible 3D shape variations. Compared with the 3D shape model approach, the 2D + 3D AAM has the advantage of computational efficiency. However, the 2D AAM requires a large number of parameters.

The advantage of employing 3D pose information is the ability to render the 3D model from new view points. Thus it is useful for 3D reconstruction. However, this group of approaches is limited to dealing with the nonlinear variations caused by face pose, and the nonrigid deformation due to facial expression is not considered. In addition, these methods often suffer from significant time complexity, impeding them for real-time applications.

(3) *Nonlinear models*: Rather than utilizing a set of linear models, Sozou et al. [25] propose a nonlinear polynomial regression point distribution model (PRPDM), which can approximate the nonlinear modes of shape variations by using polynomial functions. However, it requires human intervention to determine the degree of polynomial for each mode. Moreover, the PRPDM could only succeed in representing limited nonlinear shape variability.

Romdhani et al. [26] introduce nonlinearity into a 2D appearance model by using Kernel PCA (KPCA). A view-context-based ASM is proposed to model the shape and the pose of the face under different view points. A set of 2D shape vectors indexed with the pose angles and the corresponding gray-level appearances around feature points are used to model the multi-view face through KPCA. The nonlinearity enables the model to handle the large variations caused by face pose, but the computational complexity is prohibitive.

In summary, despite these efforts, previous technologies often focus on only one of the source of nonlinear variations either caused by face pose or by the nonrigid deformation due to facial expression, while ignoring the other. In real applications, the nonlinearity from both of face pose variation and facial expression changes should be taken into account. For example, tracking mouth and eye motion in image sequences is especially difficult, since these facial components are highly deformable, varying in shape, color, and size resulting from simultaneous variation in facial expression and head pose.

3. Facial feature tracking algorithm

3.1. Hierarchical multi-state pose-dependent facial shape model

To model 3D facial deformation and 3D head movement, we assume that a 3D facial model can be represented by a set of dominant facial feature points. The relative movements of these points characterize facial expression and face pose changes. Fig. 2 shows the layout of feature points including fiducial points and contour points.

Fiducial points are the key points on the human faces. They are located at well-defined positions such as eye corners, top points of eyebrows, and mouth corners. Fiducial points are further divided into global and local fiducial points. The global fiducial points are relatively stable with respect to facial expression change, and their movements are primarily caused by head movements. The local fiducial points are the dominant points located along the boundary of a facial component. Contour points are interpolated between the fiducial points along the boundary of a facial component. They, along with the local fiducial points, are primarily used to characterize facial expression change.

(1) *Point distribution model*: Given these facial feature points for a particular face view, a PDM can be constructed to characterize possible shape variations of human faces. Using the principle of the ASM, the PDM is constructed from a training set of face images. Facial feature points are marked on each face to outline its structure characteristics, and for each image a shape vector is used to represent the positions of feature points. All face shape vectors are aligned into a common coordinate frame by Procrustes transform [34]. Then the spatial constraints within feature points are captured by PCA [10]. A face shape vector $\mathbf{s}$ can be approximated by

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

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

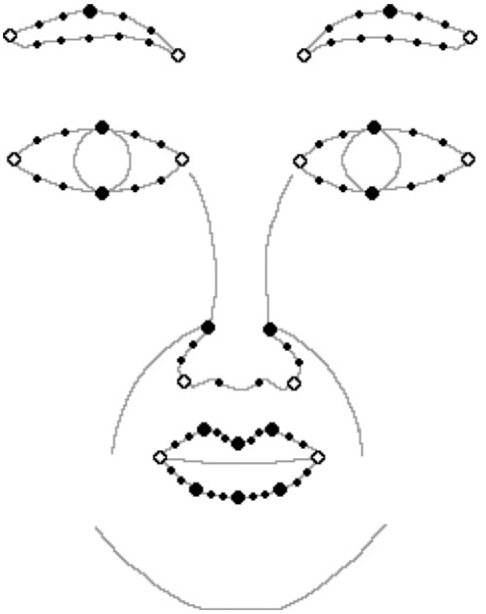

Fig. 2. Feature points in the facial model: fiducial points marked by circles (global) and big black dots (local), and contour points marked by small black dots.

The face shape can be deformed by controlling the shape parameters. By applying limits to the elements of the shape parameter vector, it is possible to ensure that the generated shape is an admissible configuration of human face. The ASM approach searches the face shape by an iterative procedure. At each iteration the algorithm seeks to match each feature point locally and then refine all feature point locations by projecting them to the PCA shape space of the entire face. Fitting the entire face shape helps mitigate the errors in individual feature matchings.

(2) *Hierarchical shape model*: In the conventional ASM, all the feature point positions are updated (or projected) simultaneously, which indicates that the interactions within feature points are simply parallel. Intuitively, human faces have a sophisticated structure, and a simple parallel mechanism may not be adequate to describe the interactions among facial feature points. For example, given the corner points of an eye, whether the eye is open or closed (or the top and bottom points of the eye) will not affect the localization of mouth or nose. This implies that the eye corners determine the overall location of the eye and provide global shape characteristics, while the top and bottom points of the eye determine the state of the eye (open or closed) and contain local structural details. Generally, facial feature points can be organized into two categories: global feature points and local feature points. The first class characterizes the global shape constraints for the entire face, while the second class captures the local structural details for individual facial components such as eyes and mouth. Based on the two-level hierarchy in facial feature points, a hierarchical formulation of statistical shape models is presented in this section.

The face shape vector $\mathbf{s}$ now could be expressed as $(\mathbf{s}_g, \mathbf{s}_l)^{\mathrm{T}}$, where $\mathbf{s}_g$ and $\mathbf{s}_l$ denote global and local feature points, respectively. The global and local fiducial points are marked by circles and large black dots as shown in Fig. 2, respectively. All the contour points (marked as small black dots) are used as local feature points in our facial model. It can be seen from the facial model that the global feature points are less influenced by local structural variations such as eye open or closed, compared to the local feature points. 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 PDM can be learned from the training data,

$$\mathbf{s}_g = \bar{\mathbf{s}}_g + \mathbf{P}_g\mathbf{b}_g, \tag{2}$$

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.

The local shape model for the ith component is denoted by a shape vector $\mathbf{s}_{g_i,l_i}$, where $\mathbf{s}_{g_i,l_i} = \{\mathbf{s}_{g_i}, \mathbf{s}_{l_i}\}$ and $\mathbf{s}_{g_i}$ and $\mathbf{s}_{l_i}$

represent the global and local feature points belonging to the ith facial component, respectively. We have

$$\mathbf{s}_{g_i,l_i} = \bar{\mathbf{s}}_{g_i,l_i} + \mathbf{P}_{g_i,l_i}\mathbf{b}_{g_i,l_i}, \tag{3}$$

where $\bar{\mathbf{s}}_{g_i,l_i}$, $\mathbf{P}_{g_i,l_i}$, and $\mathbf{b}_{g_i,l_i}$ are the corresponding mean shape vector, principal orthogonal modes of shape variation, and shape parameters for the ith facial component, respectively.

The feature search procedure in the hierarchical shape model is divided into two stages. In the first step, the positions of global feature points are matched, and shape parameters are updated iteratively using the global shape model in Eq. (2), which is the same as the search process in the ASM. In the second step, only the local feature points are matched for each facial component, and shape parameters are updated iteratively using the local shape model in Eq. (3), meanwhile positions of the global feature points remain unchanged. Therefore, in the hierarchical formulation, the global shape model and local shape models form shape constraints for the entire face and individual components, respectively. The positions of global feature points will help the localization of local feature points from each facial component. Furthermore, given the global feature points, localization of the local feature points belonging to one facial component will not affect locating the local feature points of other components. For example, given the rough positions of eyes and mouth, whether the eyes are open does not affect locating the feature points on the lips.

(3) *Multi-state local shape model*: Since it is difficult for a single-state statistical shape model to handle the nonlinear shape deformations of certain facial components such as open or closed eyes and mouth, multi-state local shape models are further introduced to address facial expression change. Specifically, for our local facial models, 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). Given this, for the ith facial component under the jth state, the shape vector becomes $\mathbf{s}_{g_i,l_{i,j}}$, and the local shape model in Eq. (3) is extended by the multi-state formulation:

$$\mathbf{s}_{g_i,l_{i,j}} = \bar{\mathbf{s}}_{g_i,l_{i,j}} + \mathbf{P}_{g_i,l_{i,j}}\mathbf{b}_{g_i,l_{i,j}}, \tag{4}$$

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 for the ith facial component at the jth state.

Hence a local shape model is built for each state of the facial component. Given the states of facial components, feature points can be searched by the two-stage procedure described in Section 3.1-(2). Therefore, the multi-state hierarchical shape model consists of a global shape model and a set of multi-state local shape models.

(4) *Pose-dependent face shape model*: The hierarchical face shape model, which we have introduced so far, basically assumes normative frontal face. The shape model (both local and global) will vary significantly, if face pose moves away from the frontal face. To compensate facial shape deformation due to face pose, we propose to estimate the 3D face pose and then use the estimated 3D pose to correct the hierarchical shape model.

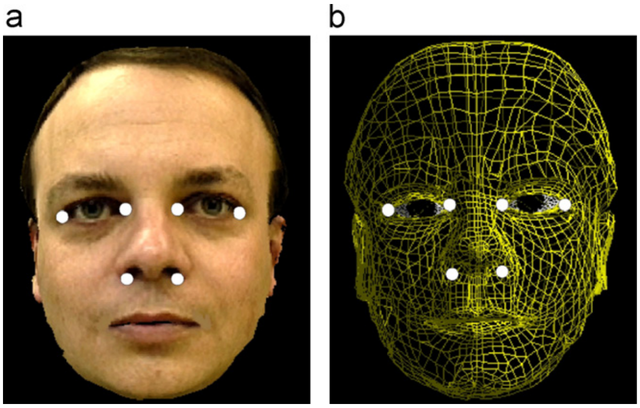

Fig. 3. A synthesized frontal face image (a) and its 3D face geometry (b) with the rigid facial feature points marked by the white dots.

Robust face pose estimation: Given detected feature points at the previous frame, the 3D face pose can be efficiently estimated. In order to minimize the effect of facial expressions, only a set of rigid feature points that are not sensitive to facial expression changes is selected to estimate the face pose. Specifically, six feature points are selected, which include the four eye corners and two fiducial points at the nose's bottom shown in Fig. 3(a).

In order to estimate the face pose, the 3D shape model composed of these six facial features has to be initialized. Currently, the coordinates $\mathbf{X}_i = (x_i, y_i, z_i)^{\mathrm{T}}$ of the six facial feature points in the 3D facial shape model are first initialized from a generic 3D face model as shown in Fig. 3(b). Due to the individual difference with the generic face model, the x and y coordinates of each facial feature point in the 3D face shape model are adjusted automatically to the specific individual based on the detected facial feature points in the initial frontal face view image. Since the depth values of the facial feature points are not available for the specific individual, the depth pattern of the generic face model is used to approximate the z_i value for each facial feature point. Our experiment results show that this method is effective and feasible in our real-time application.

Based on the personalized 3D face shape model and these six detected facial feature points in a given face image, the face pose vector $\alpha = (\sigma_{pan}, \phi_{tilt}, \kappa_{swing}, \lambda)^{\mathrm{T}}$ can be estimated accurately, where $(\sigma_{pan}, \phi_{tilt}, \kappa_{swing})$ are the three face pose angles and λ is the scale factor. Because the traditional least-square method [35] cannot handle the outliers successfully, a robust algorithm based on RANSAC [36] is employed to estimate the face pose accurately.

The pose estimation algorithm is briefly summarized as follows. The procedure starts with randomly selecting three feature points to form a triangle T_i. Under weak perspective projection model [37], each vertex (c_k, r_k) of T_i in the given image and the corresponding point (x_k, y_k) on the 3D face model are related as follows:

$$\begin{pmatrix} c_k - c_0 \\ r_k - r_0 \end{pmatrix} = \mathbf{M}_i \begin{pmatrix} x_k - x_0 \\ y_k - y_0 \end{pmatrix}, \tag{5}$$

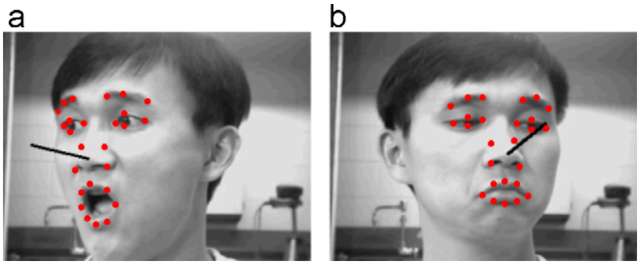

Fig. 4. The face pose estimation results under different facial expressions: face normal is represented by the dark line, and the detected facial feature points are marked by the red dots.

where $k = 1, 2$, and 3; $\mathbf{M}_i$ is the projection matrix; (c_0, r_0) and (x_0, y_0) are the centers of the triangle in the given image and the reference 3D face model, respectively. Given the three detected feature points and the corresponding points on the 3D face model, we can solve the projection matrix $\mathbf{M}_i$ for T_i. Using $\mathbf{M}_i$, the face model is projected onto the given image. A projection error e_i is then computed for all six feature points. e_i is then compared with a threshold e_0, which is determined based on the amount of outliers estimated. $\mathbf{M}_i$ is discarded, if e_i is larger than e_0. Otherwise, a weight ω_i is computed as $(e_i - e_0)^2$ for $\mathbf{M}_i$. After repeating the above procedure for each triangle formed by the six feature points, we will get a list of matrices $\mathbf{M}_i$ and their corresponding weights ω_i. From each projection matrix $\mathbf{M}_i$, a face pose vector α_i is computed uniquely after imposing some consistency constraints. Then the final face pose vector can be obtained as:

$$\alpha = \frac{\sum_{i=1}^{K} \alpha_i * \omega_i}{\sum_{i=1}^{K} \omega_i}. \tag{6}$$

Fig. 4 shows some face pose estimation results, where the face normal is perpendicular to the face plane and represented by the three estimated Euler face pose angles.

Face shape compensation: Given the estimated 3D face pose, the hierarchical model is modified accordingly. Specifically, for each frame, the mean global and local shapes are modified by projecting them to the image plane using the estimated face pose through Eq. (5). The modified mean shapes are more suitable for the current pose and provide better shape constraints for the feature search. Moreover, the projected mean shapes offer good initialization to avoid being trapped into local minima during the feature search process.

3.2. Multi-modal facial feature tracking

A multi-modal tracking approach is required to enable the state switching of facial components during the feature tracking process. 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, and the global shape model in Eq. (2) remains unchanged. In this work, the switching hypothesized measurements (SHM) model [38] is applied to dynamically estimate both the component state s_t and feature point positions of each facial component at time instant t.

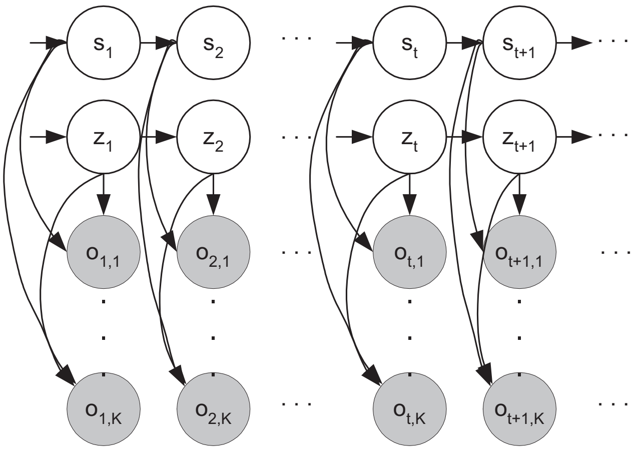

Fig. 5. Directed acyclic graph specifying conditional independence relations for switching hypothesized measurements model, where s_t is the component state, $\mathbf{z}_t$ represents the positions and velocities of facial feature points belong to each facial component, and $\mathbf{o}_{t,j},\ j \in 1, \ldots, K$ are the measurements at time t.

For feature points of a facial component, the hidden state $\mathbf{z}_t$ represents their positions and velocities at time instant t. The hidden state transition can be modeled as

$$\mathbf{z}_t = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{n}, \tag{7}$$

where $\mathbf{F}$ is the state transition matrix, and $\mathbf{n}$ represents the system perturbation with the covariance matrix $\mathbf{Q}$.

The switching state transition is modeled by a first order Markov process to encourage the temporal continuity of the component state. For the local shape model of ith component with K possible states, Fig. 5 illustrates the SHM model by a representation of dynamic Bayesian network, where the nodes s_t and $\mathbf{z}_t$ are hidden nodes, and the shaded nodes $\mathbf{o}_{t,j},\ j \in 1, \ldots, K$ are observation nodes representing the measurements at time t. Since the component state s_t is unknown, feature point positions are searched once under each hypothesized 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 the component state 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 $j \neq s_t$ 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

$$\mathbf{o}_{t,j} = \begin{cases} \mathbf{H}\mathbf{z}_t + \mathbf{v}_{t,j} & \text{if } j = s_t, \\ \mathbf{w} & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbf{H}$ is the measurement matrix; $\mathbf{v}_{t,j}$ represents the measurement uncertainty assuming as a zero mean Gaussian with the covariance matrix $\mathbf{R}_j$; 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 [39]. The state transition matrix $\mathbf{F}$ relates the state vector $\mathbf{z}_{t-1}$ at the previous time step $t-1$ to the state vector $\mathbf{z}_t$ at the current step t. The measurement matrix $\mathbf{H}$ relates the state vector $\mathbf{z}_t$ to the measurement $\mathbf{o}_{t,j}$ at current time step t under the correct state assumption, i.e. $j = s_t$. In this work, we assume the matrices $\mathbf{F}$ and $\mathbf{H}$ are stationary.

At time t, given the state transition model, measurement model, and the measurements, the facial feature tracking is performed by maximizing the posterior probability:

$$p(s_{t+1}=i, \mathbf{z}_{t+1}|\mathbf{o}_{1:t+1}) = p(s_{t+1}=i|\mathbf{o}_{1:t+1}) \times p(\mathbf{z}_{t+1}|s_{t+1}=i, \mathbf{o}_{1:t+1}), \quad (9)$$

where $\mathbf{o}_m = \{\mathbf{o}_{m,1}, \ldots, \mathbf{o}_{m,K}\}$, $1 \leqslant m \leqslant t+1$.

Let $\beta_{t,i} = p(s_t = i|\mathbf{o}_{1:t})$ with $\sum_i \beta_{t,i} = 1$, and assume that $p(\mathbf{z}_t|s_t = i, \mathbf{o}_{1:t})$ is modeled by a Gaussian distribution $N(\mathbf{z}_t; \mu_{t,i}, \mathbf{P}_{t,i})$, with the mean $\mu_{t,i}$ and covariance matrix $\mathbf{P}_{t,i}$. Then at time $t+1$, given the hypothesized measurements $\mathbf{o}_{t+1}$, the parameters $\{\beta_{t+1,i}, \mu_{t+1,i}, \mathbf{P}_{t+1,i}\}$ are updated in the SHM filtering algorithm [38] as below:

$$\beta_{t+1,i} = p(s_{t+1}=i|\mathbf{o}_{1:t+1}) = \frac{\sum_j^K \gamma_{i,j}\beta_{t,j} N(\mathbf{o}_{t+1,i}; \mathbf{H}\mu_{t+1|t,j}, \mathbf{S}_{t+1,i|j})}{\sum_i^K \sum_j^K \gamma_{i,j}\beta_{t,j} N(\mathbf{o}_{t+1,i}; \mathbf{H}\mu_{t+1|t,j}, \mathbf{S}_{t+1,i|j})}, \quad (10)$$

where $\gamma_{i,j}$ is the state transition probability:

$$\gamma_{i,j} = p(s_{t+1}=i|s_t=j) \quad (11)$$

with $\sum_i^K \gamma_{i,j} = 1$:

$$\mu_{t+1|t,j} = \mathbf{F}\mu_{t,j}, \quad (12)$$

$$\mathbf{P}_{t+1|t,j} = \mathbf{F}\mathbf{P}_{t,j}\mathbf{F}^{\mathrm{T}} + \mathbf{Q}, \quad (13)$$

$$\mathbf{S}_{t+1,i|j} = \mathbf{H}\mathbf{P}_{t+1|t,j}\mathbf{H}^{\mathrm{T}} + \mathbf{R}_i, \quad (14)$$

$$\mathbf{G}_{t+1,i|j} = \mathbf{P}_{t+1|t,j}\mathbf{H}^{\mathrm{T}}\mathbf{S}^{-1}_{t+1,i|j}, \quad (15)$$

where $\mathbf{G}$ is the gain matrix.

$$\mu_{t+1,i|j} = \mu_{t+1|t,j} + \mathbf{G}_{t+1,i|j}(\mathbf{o}_{t+1,i} - \mathbf{H}\mu_{t+1|t,j}), \quad (16)$$

$$\mathbf{P}_{t+1,i|j} = \mathbf{P}_{t+1|t,j} - \mathbf{G}_{t+1,i|j}\mathbf{H}\mathbf{P}_{t+1|t,j}, \quad (17)$$

$$\beta_{t+1,i|j} = \frac{\gamma_{i,j}\beta_{t,j} N(\mathbf{o}_{t+1,i}; \mathbf{H}\mu_{t+1|t,j}, \mathbf{S}_{t+1,i|j})}{\sum_j^K \gamma_{i,j}\beta_{t,j} N(\mathbf{o}_{t+1,i}; \mathbf{H}\mu_{t+1|t,j}, \mathbf{S}_{t+1,i|j})}, \quad (18)$$

$$\mu_{t+1,i} = \sum_j^K \beta_{t+1,i|j}\mu_{t+1,i|j}, \quad (19)$$

$$\mathbf{P}_{t+1,i} = \sum_j^K \beta_{t+1,i|j}[\mathbf{P}_{t+1,i|j} + (\mu_{t+1,i|j} - \mu_{t+1,i})(\mu_{t+1,i|j} - \mu_{t+1,i})^{\mathrm{T}}]. \quad (20)$$

Therefore, the hidden state $\mathbf{z}_{t+1}$ and the switching state s_{t+1} could be estimated as

$$\hat{s}_{t+1} = \operatorname*{argmax}_i \beta_{t+1,i}, \quad (21)$$

$$\hat{\mathbf{z}}_{t+1} = \mu_{t+1,\hat{s}_{t+1}|\hat{s}_t}. \quad (22)$$

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.

Given the multi-state hierarchical shape model, the facial feature detection and tracking algorithm performs an iterative process at time t:

(1) Project the mean shapes of global model $\bar{\mathbf{s}}_g$ and local shape models $\bar{\mathbf{s}}_{g_i,l_i}$ using the estimated face pose α_{t-1} from the previous frame.
(2) Localize the global facial feature points $\mathbf{s}_g$ individually.
(3) Update the global shape parameters to match $\mathbf{s}_g$, and apply the constraints on $\mathbf{b}_g$.
(4) Generate the global shape vector $\mathbf{s}_g$ as Eq. (2), and then return to Step 2 until convergence.
Enumerate all the possible states of the ith facial components. Under the assumption of the jth state:
(5) Localize the local feature points individually.
(6) Update the local shape parameters to match $\mathbf{s}_{g_i,l_{i,j}}$, and apply the constraints on $\mathbf{b}_{g_i,l_{i,j}}$.
(7) Generate the shape vector $\mathbf{s}_{g_i,l_{i,j}}$ as Eq. (4), and then return to Step 5 until convergence.
(8) Take the feature search results $(\mathbf{s}_{g_i}, \mathbf{s}_{l_{i,j}})^{\mathrm{T}}$ for the ith facial component under different state 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 as Eqs. (21) and (22).
(9) Estimate the 3D face pose α_t by the tracked six rigid facial feature points.

3.3. Hybrid facial feature representation

For feature detection, a hybrid feature representation, based on Gabor wavelet jets [40] and gray-level profiles [10], is utilized in this work to model the local information of fiducial points and contour points, respectively.

(1) *Wavelet-based representation*: Multi-scale and multi-orientation Gabor-wavelets [40] are employed to model local appearances around fiducial points. Gabor-wavelet-based feature representation has the psychophysical basis of human vision and achieves robust performance for expression recognition [41,42], face recognition [13], and facial feature representation [16,43] under illumination and appearance variations.

For a given pixel $\mathbf{x} = (x, y)^{\mathrm{T}}$ in a gray scale image I, a set of Gabor coefficients $J_j(\mathbf{x})$ is used to model the local appearance around the point. The coefficients $J_j(\mathbf{x})$ are resulted from convolutions of image $I(\mathbf{x})$ with the 2D Gabor wavelet kernels ψ_j, i.e.

$$J_j(\mathbf{x}) = \sum\sum I(\mathbf{x}')\psi_j(\mathbf{x}-\mathbf{x}'). \tag{23}$$

Here kernel ψ_j is a plane wave restricted by a Gaussian envelope function:

$$\psi_j(\mathbf{x}) = \frac{\mathbf{k}_j^2}{\sigma^2}\exp\left(-\frac{\mathbf{k}_j^2\mathbf{x}^2}{2\sigma^2}\right)\left[\exp(i\mathbf{k}_j\cdot\mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{24}$$

with the wave vector

$$\mathbf{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_v \cos\varphi_u \\ k_v \sin\varphi_u \end{pmatrix}, \tag{25}$$

where $k_v = 2^{-(v+1)}$ with $v = 0, 1, 2$ is the radial frequency in radians per unit length; $\varphi_u = (\pi/6)u$ with $u = 0, 1, \ldots, 5$ is the wavelet orientation in radians, rotated counter-clockwise around the origin; $j = u + 6v$; and $\mathrm{i} = \sqrt{-1}$ in this section. In this work, $\sigma = \pi$ is set for a frequency bandwidth of one octave.

Thus the set of Gabor kernels consists of three spatial frequencies and six different orientations, and 18 Gabor coefficients in the complex form are used to represent the pixel and its vicinity. Specifically, a jet vector $\mathbf{J}$ is used to denote $(J_0, J_1, \ldots, J_{17})$, where $J_j = a_j \exp(\mathrm{i}\phi_j)$, a_j and ϕ_j are the magnitude and phase of the jth Gabor coefficient. The Gabor wavelet jet vector is calculated for each marked fiducial point in training images. Given a new image, the fiducial points are searched by the sample jets from the training data. The similarity between two jet vectors is measured with the following phase-sensitive distance function:

$$D_\phi(\mathbf{J}, \mathbf{J}') = 1 - \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \mathbf{d}\cdot\mathbf{k}_j)}{\sqrt{\sum_j a_j^2 * \sum_j {a'_j}^2}}, \tag{26}$$

where jet vectors $\mathbf{J}$ and $\mathbf{J}'$ refer to two locations with relative small displacement $\mathbf{d}$. The basic idea to estimate the displacement $\mathbf{d}$ is from the Fourier shift property, i.e. a shift $\mathbf{d}$ in the spatial domain can be detected as a phase shift $\mathbf{k}\cdot\mathbf{d}$ in the frequency domain. In another word, the phase change $\Delta\phi$ is proportional to the displacement $\mathbf{d}$ along the direction of the local frequency $\mathbf{k}$. The displacement between the two locations can be approximately estimated by minimizing the phase-sensitive distance $D_\phi(\mathbf{J}, \mathbf{J}')$ in Eq. (26) as in Refs. [44,45]:

$$\mathbf{d}(\mathbf{J}, \mathbf{J}') = \begin{pmatrix} d_x \\ d_y \end{pmatrix} \approx \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \times \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} \tag{27}$$

if $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx} \neq 0$, where

$$\Phi_x = \sum_j a_j a'_j k_{jx}(\phi_j - \phi'_j),$$

$$\Gamma_{xy} = \sum_j a_j a'_j k_{jx} k_{jy},$$

and Φ_y, Γ_{xx}, Γ_{yx}, and Γ_{yy} are defined accordingly.

The phase-sensitive distance defined in Eq. (26) changes rapidly with location, which helps accurately localize fiducial points in the image. Compensated by the displacement in Eq. (27), the search of fiducial points can achieve subpixel sensitivity.

(2) *Profile-based representation*: Gray-level profiles (gradients of pixel intensity) along the normal direction of the object boundary are used to represent contour points [10]. The Mahalanobis distance function is used to search these feature points. For the lth contour point,

$$D_M(\mathbf{g}, \bar{\mathbf{g}}_l) = (\mathbf{g} - \bar{\mathbf{g}}_l)^{\mathrm{T}}\mathbf{C}_l^{-1}(\mathbf{g} - \bar{\mathbf{g}}_l), \tag{28}$$

where $\mathbf{g}$ is a gray-level profile in the given image; $\bar{\mathbf{g}}_l$ is the mean profile of the lth contour point computed from the training data; and $\mathbf{C}_l$ is the corresponding covariance matrix obtained by training.

The profile-based representation is computationally efficient, and thus leads to fast convergence in the feature search process. However, the gray-level gradients are not sufficient to identify all the facial feature points. For example, the profile of the mouth bottom point may not be distinctive due to the shadow or beard below the lower lip. On the other hand, the magnitudes and phases in wavelet-based representation provide rich information of local appearances, and therefore lead to accurate feature point localization, but with relatively high computation complexity. To balance the searching effectiveness and computational efficiency, in this work the fiducial points are modeled by Gabor wavelet jets, and the contour points, which are relatively less important to the face shape deformation, are represented by gray-level profiles. 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.

(3) *Feature detection*: Given the hybrid representation for each feature point, we can perform feature detection. Feature detection starts with eye detection. An accurate and robust eye detector is desirable to help estimate the position, size, and orientation of the face region, and thus improve the shape constraints for the feature search. In this work, a boosted eye detection algorithm is employed based on recursive nonparametric discriminant analysis (RNDA) features proposed in [46,47]. For eye detection, the RNDA features provide better accuracy than Harr features [48], since they are not constrained with rectangle-like shape. The features are sequentially learned and combined with Adaboost to form an eye detector. To improve speed, a cascade structure is applied. The eye detector is trained on thousands of eye images and more non-eye images. The resulting eye detector classifier uses less than 100 features.

The eye localization follows a hierarchical principle: first a face is detected, then the eyes are located inside the detected face. An overall 94.5% eye detection rate is achieved with 2.67% average normalized error (the pixel error normalized by the distance between two eyes) on FRGC 1.0 database [49]. Given the knowledge of eye centers, the face region is normalized and scaled into a 64 × 64 image, such that the eyes are nearly fixed in same positions in each frame.

Given the normalized face region, the other facial feature points as illustrated in Fig. 2 are detected based on the hybrid feature representations: the contour points are searched by minimizing the Mahalanobis distance defined in Eq. (28); and the fiducial points are detected by minimizing the phase-sensitive distance function in Eq. (26).

4. Experimental results

Twenty-six fiducial points and 56 contour points are used in our facial model (see Fig. 2). 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 containing 200 persons from different races, ages, face poses, and facial expressions. Both feature point positions and facial component states are manually labeled in each training image. For ASM analysis, the principal orthogonal modes in the shape models stand for 95% of the shape variation. The test sequences consist of 10 sequences, each of which consists 100 frames. The test sequences contain six subjects, which are not included in training data. The test sequences are 24-bit color images collected by a USB web camera under real-world conditions with 320 × 240 image resolution. The system can reliably detect and track the face in a range of 0.25–1.0 m from the camera. Since the face region is normalized and scaled based on the detected eye positions, the system is invariant to the scale change. Our C + + program can process about seven frames per second on a Pentium 4 2.8 GHz PC.

4.1. Experiments on facial feature tracking

Figs. 6–8 exhibit the results of facial feature tracking under pose variations and face deformations due to varying facial expressions. To make the results clear, only the fiducial points are shown, since they are more representative than the contour points. Fig. 6 shows the results of the proposed method and the results without using the pose modified mean shapes. Compared to the results in Fig. 6(b), the feature points are more robustly tracked in Fig. 6(a) under large pose variations, which demonstrates that the projection of mean shapes through face pose estimation helps improve shape constraints in the feature search process. Fig. 7 shows the results of the proposed method and the results without using the multi-state local shape models (i.e. a single-state local shape model is used for each facial component). It can be seen from Fig. 7 that the multi-state models substantially improve the robustness of facial feature tracking, especially when eyes and mouth are open or closed (e.g. mouth in the first and fourth images, and eyes in the second and third images). This demonstrates that the state switching in local shape models helps in dealing with nonlinear shape deformations of facial components.

Fig. 8 shows the results of the proposed method and the results without using the hierarchical shape model (i.e. all the feature points are simultaneously updated in the feature search process). Compared to the results in Fig. 8(b), the feature points are more accurately tracked in Fig. 8(a) (e.g. mouth in the first image, right eye in the second and third images, and left eyebrow in the fourth image) using the hierarchical shape model, since the two-level hierarchical facial shape model provides

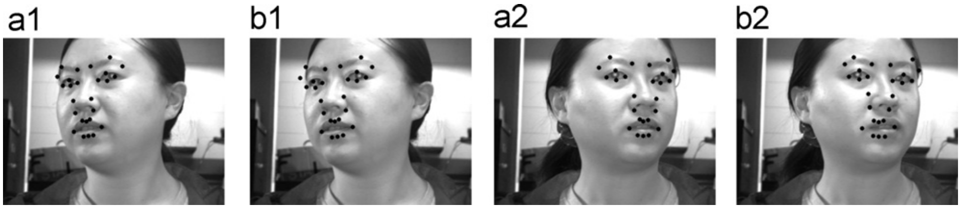

Fig. 6. Feature tracking results: (a) by the proposed method and (b) by the proposed method without using modified mean shapes.

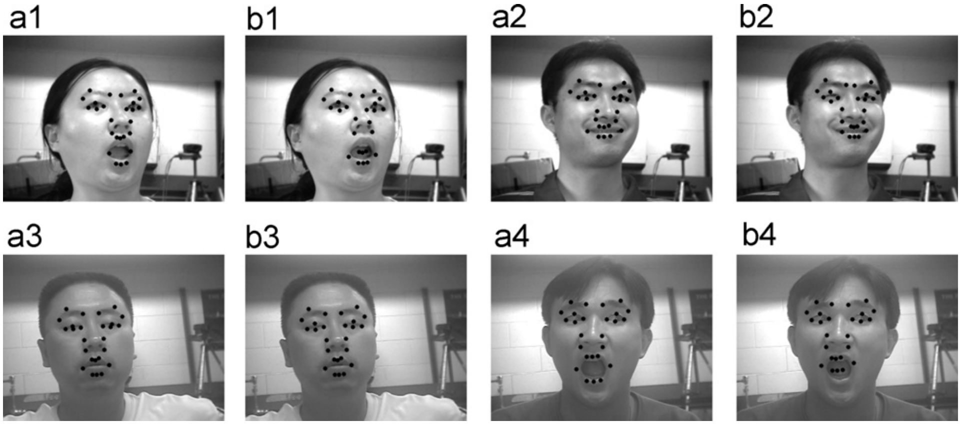

Fig. 7. Feature tracking results: (a) by the proposed method and (b) by the proposed method without using the multi-state local shape models.

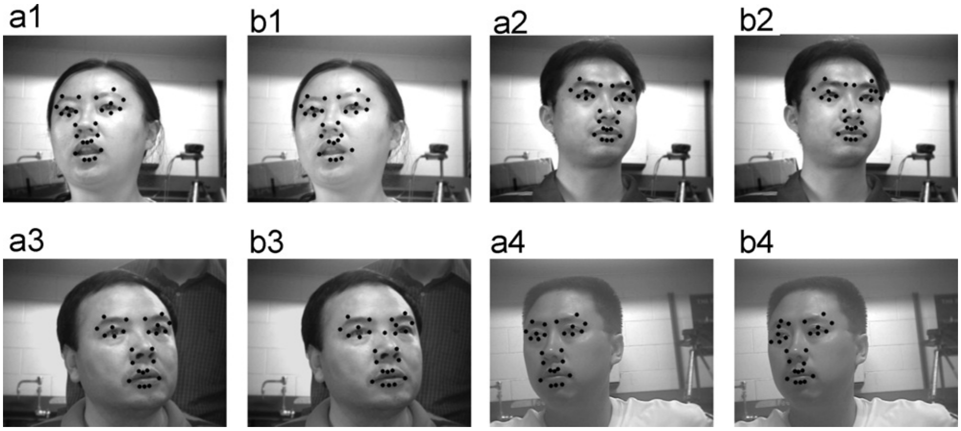

Fig. 8. Feature tracking results: (a) by the proposed method and (b) by the proposed method without using the hierarchical shape model.

Fig. 9. A set of testing images from an image sequence, where the face undergoes the face pose change and facial expression change simultaneously.

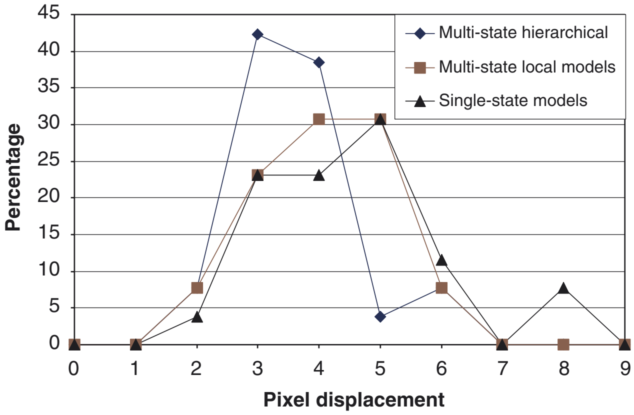

Fig. 10. Error distribution of feature tracking. Diamonds: results of the proposed method. Triangles: results of the proposed method without the multi-state local shape models. Squares: results of the proposed method without the hierarchical shape model.

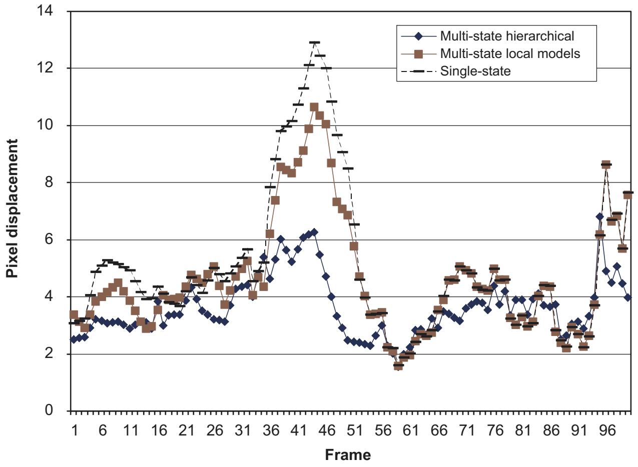

Fig. 11. Error evolution of feature tracking for an image sequence. Diamonds: results of the proposed method. Bars: results of the proposed method without the multi-state local shape models. Squares: results of the proposed method without the hierarchical shape model.

a relatively sophisticated structure to describe the interactions among feature points.

4.2. *Quantitative evaluation*

The results of facial feature tracking are evaluated quantitatively besides visual comparison. Fiducial and contour feature points are manually marked in 1000 images from the 10 test sequences for comparison under different face pose and facial expressions. Fig. 9 show a set of testing images from an image sequence, where the face undergoes the face pose change and facial expression change simultaneously. For each feature point, the displacement (in pixels) between the estimated position and the corresponding labeled position is computed as the feature tracking error. Fig. 10 shows error distribution of the feature tracking results of different methods. Compared to the feature tracking result using single-state local model and using multi-state local models without the hierarchical models, the use of the multi-state hierarchical shape model averagely reduces the feature tracking error by 24% and 13%, respectively. Besides the comparison of the average pixel displacement, Fig. 11 illustrates the evolution of the error over time for one image sequence, where the face undergoes both significant

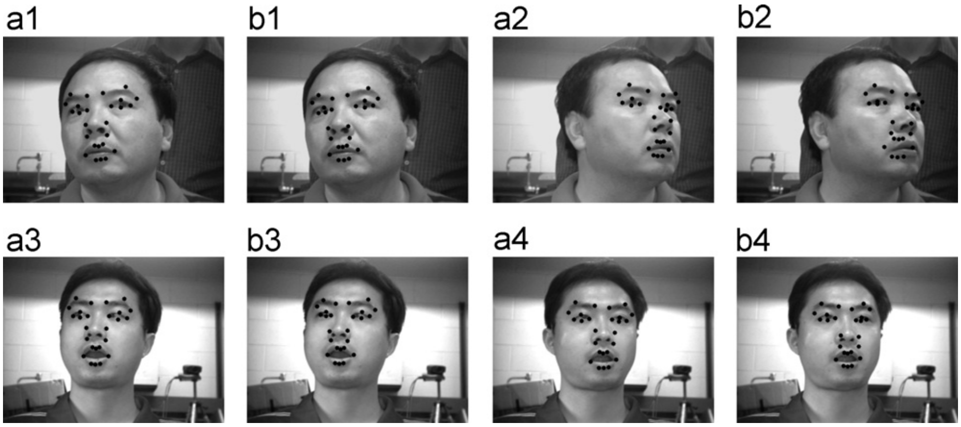

Fig. 12. Feature tracking results: (a) proposed method and (b) mixture Gaussian model.

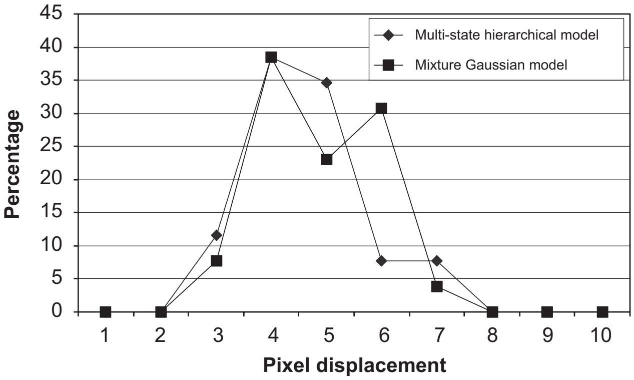

Fig. 13. Error distribution of feature tracking under face pose change. Diamonds: results of the proposed method. Squares: results of the mixture Gaussian method.

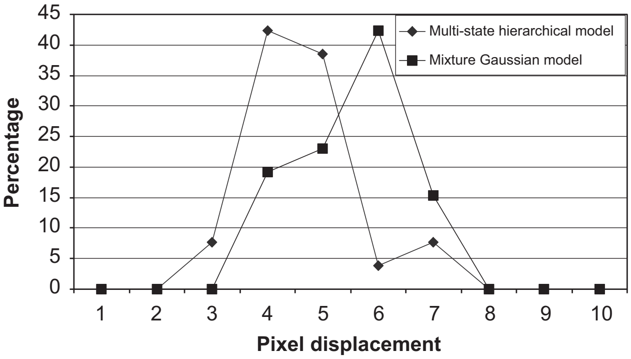

Fig. 14. Error distribution of feature tracking under both of face pose and facial expression change. Diamonds: results of the proposed method. Squares: results of the mixture Gaussian method.

facial expression and large face pose change simultaneously from frames 35 to 56. By employing the multi-state pose-dependent hierarchical shape model, the proposed method substantially improves the robustness of facial feature tracking under simultaneous facial expression and face pose variation.

4.3. Comparison with mixture of Gaussian model

Fig. 12 compares the feature tracking results of the proposed method with the results of the mixture Gaussian method by Cootes et al. [20]. It can be seen that the proposed multi-state pose-dependent hierarchical method outperforms mixture Gaussian approach under pose variations (e.g. the left eye and eyebrow in the first and third images, the mouth in the second image, and the right eye, right eyebrow and nose in the fourth image). Fig. 13 shows the error distribution of the feature tracking results using the proposed method and the mixture Gaussian method, respectively, under face pose change only. Furthermore, Fig. 14 compares the facial feature tracking results using the proposed method and the mixture Gaussian method under simultaneous facial expression and face pose change. Compared to the mixture Gaussian method, the proposed method reduces the feature tracking error by 10% and 23%, respectively, under only face pose change and the combination of facial expression and face pose change. Therefore, the proposed method is more appropriate to handle the real-world conditions, where the facial shape change due to both of facial expression and face pose. In addition, the proposed method is more efficient than the mixture Gaussian method, which runs at about 5.8 frame/second under same condition.

5. Conclusion

In this paper, a multi-state pose-dependent hierarchical face shape model is successfully developed to improve the accuracy and robustness of facial feature tracking under simultaneous pose variations and face deformations. The model allows to simultaneously characterize the global shape constraints and the local structural details of human faces. Shape constraints for the feature search are significantly improved by modifying mean shapes through robust face pose estimation. In addition, Gabor wavelet jets and gray-level profiles are integrated for effective and efficient feature representation. Feature point positions are dynamically estimated with multi-state local shape models using a multi-modal tracking approach. Experimental results demonstrate that the proposed method significantly reduces the feature tracking error, compared to the classical feature tracking methods.

In the current work, we ignore the relationships between different facial components by decoupling the face into different local models, because those relationships are complex, dynamic, and very uncertain. Moreover, incorrect modeling of such relationships will lead to the failure in detection of facial feature. In the future, we would like to combine the relationships between different facial components into the facial feature tracking by exploiting spatial–temporal relationships among different action units.

Acknowledgment

This work was supported in part by a grant from DARPA/ONR under the contract number N00014-03-1-1003.

References

[1] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, Proceedings of International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[2] J. Shi, C. Tomasi, Good features to track, Proceedings of CVPR94, 1994, pp. 593–600.

[3] F. Bourel, C. Chibelushi, A. Low, Robust facial feature tracking, Proceedings of 11th British Machine Vision Conference, vol. 1, 2000, pp. 232–241.

[4] C. Tomasi, T. Kanade, Detection and tracking of point features, Carnegie Mellon University, Technical Report CMU-CS-91-132.

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

[6] L. Torresani, C. Bregler, Space-time tracking, Proceedings of ECCV02, vol. 1, 2002, pp. 801–812.

[7] Z. Zhu, Q. Ji, K. Fujimura, K. Lee, Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination, Proceedings of ICPR02, vol. 4, 2002, pp. 318–321.

[8] M. Kass, A. WItkin, D. Terzopoulos, Snakes: active contour models, Int. J. Comput. Vision 1 (4) (1988) 321–331.

[9] A. Yuille, P. Haallinan, D.S. Cohen, Feature extraction from faces using deformable templates, Int. J. Comput. Vision 8 (2) (1992) 99–111.

[10] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Comput. Vision Image Understanding 61 (1) (1995) 38–59.

[11] T.F. Cootes, G.J. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.

[12] X.W. Hou, S.Z. Li, H.J. Zhang, Q.S. Cheng, Direct appearance models, Proceedings of CVPR01, vol. 1, 2001, pp. 828–833.

[13] L. Wiskott, J.M. Fellous, N. Krüger, C.V. der Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 775–779.

[14] M.J. Jones, T. Poggio, Multi-dimensional morphable models: a framework for representing and matching object classes, Int. J. Comput. Vision 29 (1998) 107–131.

[15] S. Sclaroff, J. Isidoro, Active blobs: region-based, deformable appearance models, Comput. Vision Image Understanding 89 (2–3) (2003) 197–225.

[16] S.J. McKenna, S. Gong, R.P. Würtz, J. Tanner, D. Banin, Tracking facial feature points with Gabor wavelets and shape models, Proceedings of International Conference on Audio- and Video-based Biometric Person Authentication, 1997, pp. 35–42.

[17] M. Rogers, J. Graham, Robust active shape model search, Proceedings of ECCV, vol. 4, 2002, pp. 517–530.

[18] T.F. Cootes, G.V. Wheeler, K.N. Walker, C.J. Taylor, View-based active appearance models, Image Vision Comput. 20 (9) (2002) 657–664.

[19] T. Heap, D. Hogg, Wormholes in shape space: tracking through discontinuous changes in shape, Proceedings of ICCV98, 1998, pp. 344–349.

[20] T.F. Cootes, C.J. Taylor, A mixture model for representing shape variation, Image Vision Comput. 17 (8) (1999) 567–573.

[21] C.M. Christoudias, T. Darrell, On modelling nonlinear shape-and-texture appearance manifolds, Proceedings of CVPR05, vol. 2, 2005, pp. 1067–1074.

[22] Y. Li, S. Gong, H. Liddell, Modelling faces dynamically across views and over time, Proceedings of ICCV01, vol. 1, 2001, pp. 554–559.

[23] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, Siggraph 1999, Computer Graphics Proceedings, 1999, pp. 187–194.

[24] J. Xiao, S. Baker, I. Matthews, T. Kanade, Real-time combined 2D + 3D active appearance models, Proceedings of CVPR04, vol. 2, 2004, pp. 535–542.

[25] P.D. Sozou, T.F. Cootes, C.J. Taylor, E. di Mauro, Non-linear generalization of point distribution models using polynomial regression, Image Vision Comput. 13 (5) (1995) 451–457.

[26] S. Romdhani, S. Gong, A. Psarrou, Multi-view nonlinear active shape model using kernel PCA, Proceedings of BMVC, 1999, pp. 483–492.

[27] S. Yan, X. Hou, S.Z. Li, H. Zhang, Q. Cheng, Face alignment using view-based direct appearance models, Special issue on facial image processing, analysis and synthesis, Int. J. Imaging Syst. Technol. 13 (1) (2003) 106–112.

[28] K. Grauman, T. Darrell, Fast contour matching using approximate earth movers's distance, Proceedings of CVPR04, vol. 1, 2004, pp. 220–227.

[29] Y. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. on Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.

[30] A. Yilmaz, K. Shafique, M. Shah, Estimation of rigid and non-rigid facial motion using anatomical face model, Proceedings of ICPR02, vol. 1, 2002, pp. 377–380.

[31] H. Tao, T.S. Huang, Visual estimation and compression of facial motion parameters: elements of a 3D model-based video coding system, Int. J. Comput. Vision 50 (2) (2002) 111–125.

[32] S.K. Goldenstein, C. Vogler, D. Metaxas, Statistical cue integration in DAG deformable models, IEEE Trans. Pattern Anal. Mach. Intell. 25 (7) (2003) 801–813.

[33] J. Xiao, J. Chai, T. Kanade, A closed-form solution to non-rigid shape and motion recovery, Int. J. Comput. Vision 67 (2) (2006) 233–246.

[34] I.L. Dryden, K.V. Mardia, Statistical Shape Analysis, Wiley, Chichester, 1998.

[35] S.H. Or, W.S. Luk, K.H. Wong, I. King, An efficient iterative pose estimation algorithm, Image Vision Comput. 16 (5) (1998) 353–362.

[36] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.

[37] E. Trucco, A. Verri, Introductory Techniques for 3-D Computer Vision, Prentice-Hall, Englewood Cliffs, 1998.

[38] Y. Wang, T. Tan, K.-F. Loe, Joint region tracking with switching hypothesized measurements, Proceedings of ICCV03, vol. 1, 2003, pp. 75–82.

[39] H. Gu, Q. Ji, Z. Zhu, Active facial tracking for fatigue detection, Proceedings of Sixth IEEE Workshop on Applications of Computer Vision, 2002, pp. 137–142.

[40] J. Daugman, Complete discrete 2D Gabor transforms by neural networks for image analysis and compression, IEEE Trans. ASSP 36 (7) (1988) 1169–1179.

[41] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron, Proceedings of FGR98, 1998, pp. 454–459.

[42] Y. Tian, T. Kanade, J.F. Cohn, Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity, Proceedings of FGR02, 2002, pp. 218–223.

[43] F. Jiao, S.Z. Li, H.Y. Shum, D. Schuurmans, Face alignment using statistical models and wavelet features, Proceedings of CVPR03, vol. 1, 2003, pp. 321–327.

[44] D.J. Fleet, A.D. Jepson, Computation of component image velocity from local phase information, Int. J. Comput. Vision 5 (1) (1990) 77–104.

[45] W.M. Theimer, Phase-based binocular vergence control and depth reconstruction using active vision, CVGIP: Image Understanding 60 (3) (1994) 343–358.

[46] P. Wang, Q. Ji, Learning discriminant features for multi-view face and eye detection, Proceedings of CVPR05, vol. 1, 2005, pp. 373–379.

[47] P. Wang, M.B. Green, Q. Ji, J. Wayman, Automatic eye detection and its validation, IEEE Workshop on Face Recognition Grand Challenge Experiments (with CVPR05), vol. 3, 2005, pp. 164–164.

[48] P. Viola, M. Jones, Robust real-time object detection, Int. J. Comput. Vision 57 (2) (2004) 137–154.

[49] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, Proceedings of CVPR05, vol. 1, 2005, pp. 947–954.

About the Author—YAN TONG is currently pursuing her Ph.D. at Rensselaer Polytechnic Institute, Troy, NY. She received the B.S. degree from Zhejiang University, PR China in 1997, and M.S. degree from University of Nevada, Reno in 2004. Current research interest focuses on computer vision, pattern recognition, and human computer interaction.

About the Author—YANG WANG received his Ph.D. degree in computer science from National University of Singapore in 2004. He is currently a researcher in National ICT Australia (NICTA)'s Multimedia and Video Communication (MVC) research project. Prior to joining NICTA, he was a research fellow at Nanyang Technological University, Singapore. He also worked at Institute for Infocomm Research, Singapore and Rensselaer Polytechnic Institute, USA as a research scholar. His research interests include video analysis, sensor networks, pattern classification, multimedia processing, medical imaging, human–computer interaction, and computer vision.

About the Author—ZHIWEI ZHU received his Ph.D. degree from the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute in December 2005. He obtained a M.S. degree from Department of Computer Science, University of Nevada, Reno in August 2002, and a B.S. degree in the Department of Computer Science, University of Science and Technology of Beijing, China in July 2000. He is currently a Member of Technical Staff in the Computer Vision Laboratory of Sarnoff Corporation in Princeton, NJ. His research interests are in computer vision, pattern recognition, image processing and human–computer interaction.

About the Author—QIANG JI received his Ph.D. degree in electrical engineering from the University of Washington in 1998. He is currently an associate Professor with the Department of Electrical, Computer, and Systems engineering at Rensselaer Polytechnic Institute (RPI). Prior to joining RPI in 2001, he was an assistant professor with Department of Computer Science, University of Nevada at Reno. He also held research and visiting positions with Carnegie Mellon University, Western Research Company, and the US Air Force Research Laboratory.

Dr. Ji's research interests are in computer vision, probabilistic reasoning with Bayesian networks for decision making and information fusion, human-computer interaction, pattern recognition, and robotics. He has published over 100 papers in peer-reviewed journals and conferences. His research has been funded by local and federal government agencies including NSF, NIH, AFOSR, ONR, DARPA, and ARO and by private companies including Boeing and Honda. Dr. Ji is a senior member of the IEEE.