Machine Vision and Applications (2005) 16: 105–115
Digital Object Identifier (DOI) 10.1007/s00138-004-0161-6

Machine Vision and
Applications

Information extraction from image sequences of real-world facial expressions

Haisong Gu[1], **Qiang Ji**[2]

[1] Department of Computer Science, University of Nevada-Reno, Reno, NV 89557, USA
[2] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Received: 16 August 2003 / Accepted: 20 September 2004
Published online: 20 December 2004 – © Springer-Verlag 2004

Abstract. Information extraction of facial expressions deals with facial-feature detection, feature tracking, and capture of the spatiotemporal relationships among features. It is a fundamental task in facial expression analysis and will ultimately determine the performance of expression recognition. For a real-world facial expression sequence, there are three challenges: (1) detection failure of some or all facial features due to changes in illumination and rapid head movement; (2) nonrigid object tracking resulting from facial expression change; and (3) feature occlusion due to out-of-plane head rotation. In this paper, a new approach is proposed to tackle these challenges. First, we use an active infrared (IR) illumination to reliably detect pupils under variable lighting conditions and head orientations. The pupil positions are then used to guide the entire information-extraction process. The simultaneous use of a global head motion constraint and Kalman filtering can robustly track individual facial features even in condition of rapid head motion and significant expression change. To handle feature occlusion, we propose a warping-based reliability propagation method. The reliable neighbor features and the spatial semantics among these features are used to detect and infer occluded features through an interframe warping transformation. Experimental results show that accurate information extraction can be achieved for video sequences with real-world facial expressions.

Keywords: Facial feature tracking – Information extraction – Facial expression – Warping – Reliability propagation

1 Introduction

Automatic recognition (or analysis) of facial expression consists of three basic steps [34]: face detection, information extraction of facial expression, and expression classification. For expression recognition from a video sequence, the purpose of the information extraction is to detect facial features, track them, and capture the spatiotemporal relationships among these features. The facial information obtained in this step will eventually determine the performance of expression classification. This has been an area of active research.

Broadly speaking, we can divide the existing works in the field into general-purpose and face-specific approaches. Among the general-purpose approaches are moving-point-correspondence methods, patch-correlation methods, and optical-flow methods.

For the moving-point-correspondence methods [9,18,36–38], the common assumptions are smooth motion, limited speed, and no (or minimum) occlusions. Each method uses a smoothness-based cost function. The cost function evaluates the local deviation from smoothness and penalizes changes in both direction and magnitude for the velocity vector. The strong assumptions make them inapplicable to facial-feature tracking. Another general-purpose technique is the patch-correlation methods [1,8,25,39,40,42,54,55], which characterize each feature with a template. Feature-point detection and tracking become a process of matching features with the template. Feature detection entails seeking the position of the normalized cross-correlation peak between a template and an image (or an image region) to locate the best match. This method is rather sensitive to illumination and object-pose variation. Finally, optical-flow-based methods [31,43,51] compute a motion field for each pixel (some features) in an image. The corresponding pixels or features in the next frame can be determined based on the motion field vectors. Optical-flow extraction often assumes image-intensity constancy for corresponding pixels, which may not be the case for facial features since the intensities may change due to illumination or face-pose change. All general-purpose methods apply to a rigid body while the face is nonrigid.

Compared with the general-purpose feature-tracking techniques, the face-specific methods show higher efficiency because they exploit knowledge of the facial domain, such as 2D/3D geometric shapes, color, and brightness. Using face models is an efficient way to conquer the variability of conditions in a long face sequence. Based on *SNAKES* [22] or deformable templates [49], several model-based methods have been proposed [10,13–15,24,27,47,50,53]. Malciu and Preteux [27] proposed a typical model-based technique for facial-feature tracking. First, a deformable template is specified by a parameterized geometry, an internal energy function, and an

Correspondence to: Qiang Ji (e-mail: qji@ecse.rpi.edu)

external energy function. Then a matching procedure is conducted to search for a 2D nonrigid transformation so as to yield an optimal registration of the template model T with the reference frame into the next frame. By combining a general-purpose method [42] with specific domain knowledge, Bourel [7] presents an approach to robust facial-feature tracking. However, this method is entirely based on the visible nostrils, which are often occluded in real-world facial expressions. The Gabor wavelet method has been used for static face analysis and recognition [28,52]. Maurer et al. [29] also used Gabor wavelet jets to track facial features on a face rotating in depth. However, each point is treated independently with an equal confidence level, and no global shape constraint is imposed. The method makes it easy for features to get lost due to rapid head motion and occlusion. McKenna [30] used Gabor wavelet jets to extract each facial point and use a single point distribution model (PDM) as a global constraint to correct the missing points. Cootes et al. [64,66] proposed to perform facial-feature detection based on combining the active shape model (ASM) and active appearance model (AAM). Shape and texture are combined in PCA space. More recently, they proposed to combine AAM with the Reinforcement of Feature Responses (PRFR) model [65] or with Adaboost training [63] for more accurate facial-feature localization.

For real-world facial expressions, out-of-plane head rotation often occurs, which is the main cause of feature occlusion. The self-occlusion has recently attracted attention for dealing with spontaneous facial expression [33,44]. Moriyama et al. [33] and Torresani [44] proposed, respectively, a 3D face model method and a factorization-based method to recover the occluded facial feature. In the factorization-based method [44,45], the object-structure and camera-motion parameters are first recovered from feature points tracked in advance. Then the motion and structure are propagated to recover the missing points. These methods, however, require initial feature detection in advance to recover the structure. Also, they are not suitable for real-time implementation because of the requirement of initial structure recovery and iterative procedure. We believe that without the feature identification, it is very hard to efficiently infer feature occlusion, which is basically the local behavior.

In summary, the existing methods share some common assumptions: frontal-view pose, constant illumination, and minimum out-of-plane head motion. Unfortunately, these assumptions are not realistic. Facial-feature detection and tracking for real-world facial expressions should consider at least two critical issues: (1) detecting and tracking important facial features in a variety of lighting conditions and under rapid head movements and (2) handling self-occlusion of features due to frequent out-of-plane head rotation. In this paper, we propose a domain-specific approach. Our method is based on the integration of several practical techniques. We developed an IR-based sensing system to reliably detect pupil positions under variable lighting conditions and rapid head motion. These pupil positions provide strong constraints on the detection and tracking of other facial features. The Kalman filtering is combined with the constraints from pupil motion to predict the location of each feature in the next frame. Gabor wavelets are used to characterize each facial feature. The Gabor wavelet coefficients are updated in each frame to adaptively represent the feature profile due to facial expressions. To handle facial-feature occlusion from face pose, a warping-based reliability propagation method is proposed. The idea is to specify a warp that maps a source image into the destination image. The reliable feature points within each facial region can be used systematically to verify the newly detected feature based on the interframe warping transformation.

2 IR active facial sensing

2.1 Hardware setup

To reliably detect and track eyes, we developed an active facial sensing system. The system consists of two sets of IR LEDs, distributed evenly and symmetrically along the circumference of two coplanar concentric rings as shown in Fig. 1a. The center of both rings coincides with the camera optical axis. We use near infrared (NIR) LEDs with a nominal wavelength of 880 nm.

According to the original patent from Hutchinson [62], a bright pupil can be obtained if the eyes are illuminated with a NIR illuminator beaming light along the camera optical axis at a certain wavelength. At the NIR wavelength, pupils reflect almost all IR light they receive along the path back to the camera, producing the bright-pupil effect, very much similar to the red-eye effect in photography. If illuminated by the IR light off the camera optical axis , the pupils appear dark since the reflected light will not enter the camera lens. This produces the so-called dark-pupil effect.

For our IR illuminator, a bright-pupil image is produced when the inner ring of IR LEDs is turned on, and a dark-pupil image is produced when the outer ring is turned on. A circuitry was developed to synchronize the inner and outer LED rings with, respectively, the even and odd fields of the interlaced image, producing, respectively, the bright-pupil and dark-pupil image on the even and odd fields of a frame, as shown in Fig. 1b, c. Pupil detection can be robustly achieved from the difference image as a result of subtracting the odd field from the even field. Details on the IR illuminator and on eye detection may be found in [19].

The active sensing system allows us to accurately detect and track eyes under variable lighting conditions, for different face poses and for different individuals [54]. The detected eyes

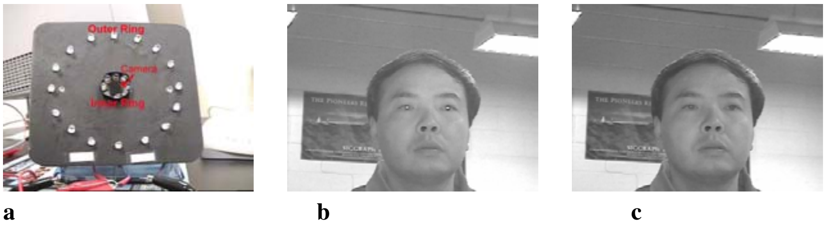

Fig. 1. a IR camera with active IR illuminator. **b** Bright pupils in an even field image. **c** Dark pupils in an odd field image

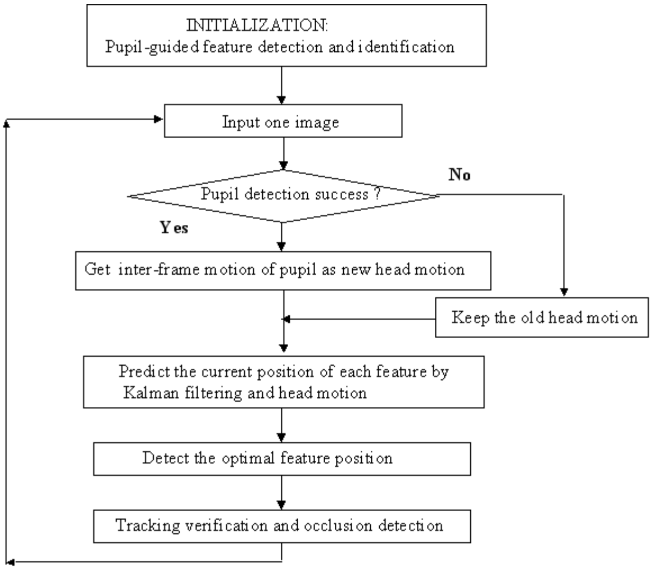
INITIALIZATION:
Pupil-guided feature detection and identification
Input one image
Pupil detection success ?
No
Yes
Get inter-frame motion of pupil as new head motion
Keep the old head motion
Predict the current position of each feature by Kalman filtering and head motion
Detect the optimal feature position
Tracking verification and occlusion detection

Fig. 2. Process flow of proposed facial-feature detection and tracking method

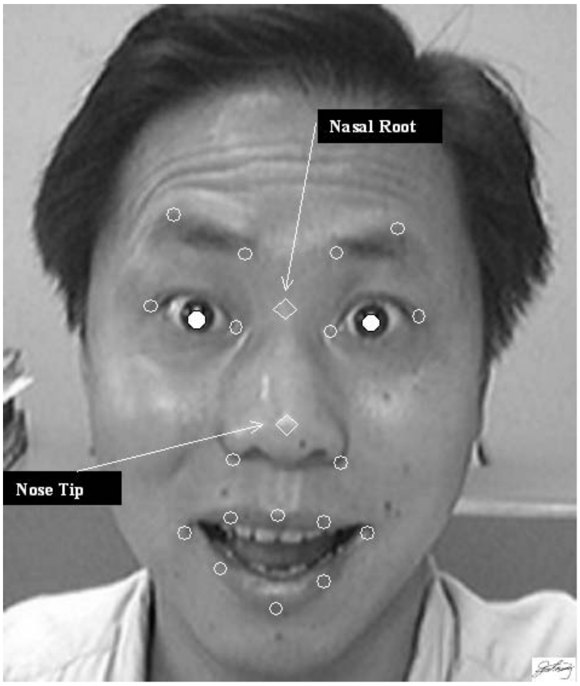

Fig. 3. Facial fiducial feature points

will help greatly in the detection and tracking of other facial features using conventional methods.

Figure 2 outlines the process flow of our facial-feature detection and tracking method. At the first frame, based on the detected pupil positions and generic face model, other facial features are detected and identified. After initialization, a Kalman tracker is used to track each individual facial feature. To improve the temporal prediction of feature position at the next frame, the Kalman prediction is linearly combined with head motion estimated by the detected pupils. The final feature position will be determined by a Gabor wavelet similarity evaluation. Finally, the individually tracked features will be verified and modified by the spatial relationships among the extracted features. Occluded features will be detected and recovered with a warping-transformation-based reliability propagation.

3 Feature extraction and tracking

3.1 Feature-based facial description

The analysis approaches to facial expression can be divided into two basic categories: appearance-based methods (such as EigenFace [46], SVM [11, 17]) and feature-based methods. So far, many research groups have reported that the feature-based approach, especially with the Gabor wavelet, has a solid psychophysical basis in human vision and can achieve good performance [12, 23, 52, 56].

The facial features around the eyes and mouth represent primary spatial patterns to compose a facial expression display. Generally, these patterns, with their spatiotemporal changes and synchronization relationships, can describe most facial expressions. On the other hand, for a specific application, such as drowsy-driver monitoring, there are only limited facial expression displays; the facial-feature points around the eyes and mouth contain enough information to capture all these related expression displays. So here, as shown in Fig. 3, we use 22 fiducial features around the eyes and mouth as the descriptors of facial expressions.

The multiscale and multiorientation Gabor wavelet is a very powerful method for describing the local property of each feature point. In order to efficiently conduct the Gabor wavelet transformation, the original image is first normalized into a 128×128 image. The 2D Gabor kernels used are as follows:

$$\Psi(\mathbf{k}, \overrightarrow{x}) = \frac{\mathbf{k}^2}{\sigma^2} e^{-\frac{\mathbf{k}^2 \overrightarrow{x}^2}{2\sigma^2}} \left(e^{i\mathbf{k}\times\overrightarrow{x}} - e^{-\frac{\sigma^2}{2}}\right), \tag{1}$$

where σ is set at π. The set of Gabor kernels consists of three spatial frequencies (with wavenumber k: $\pi/2, \pi/4, \pi/8$) and six distinct orientations from 0° to 150° in 30° intervals.

For each pixel $(\overrightarrow{x})$, a set $\Gamma(\overrightarrow{x})$ of 18 Gabor coefficients obtained by convolution with the Gabor kernels can be used to present the intensity profile of each point:

$$\begin{aligned}\Gamma(\overrightarrow{x}) &= (m_1 e^{i\phi_1}, m_2 e^{i\phi_2}, ..., m_{18} e^{i\phi_{18}})^T \qquad (2)\\ &= \int I(\overrightarrow{x}')\Psi[\mathbf{k}, (\overrightarrow{x} - \overrightarrow{x}')]d\overrightarrow{x}',\end{aligned}$$

$$\gamma(\overrightarrow{x}) = \int I(\overrightarrow{x}')\Psi[\mathbf{k}, (\overrightarrow{x} - \overrightarrow{x}')]d\overrightarrow{x}'. \tag{3}$$

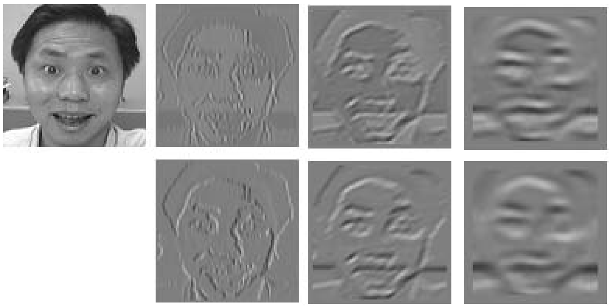

Fig. 4. Convolution results of one image with three Gabor kernels

Figure 4 shows an example of one image and its convolution results with three Gabor kernels. The upper row presents the real component images and the bottom row the imagery component images.

In our work, the coefficient set is used not only to detect and identify each facial feature in the initial frame but also to conduct tracking process as the template of each feature.

3.2 Kalman filter with pupil constraints

The face is a typical nonrigid object. The feature profile and the relationships among features can change independently according to the facial expression displays. Tracking for these features involves three basic problems: (1) feature loss due to rapid head motion, (2) random feature jumping due to noise, and (3) feature self-occlusion as a result of head turning. For the first two problems, we develop a pupil-guided Kalman filtering solution. It consists of feature prediction and feature detection. We will address the self-occlusion problem in Sect. 4.

3.2.1 Feature prediction

The pupil positions from active sensing provide reliable information that indicates roughly where the face is located and how the head changes globally. This kind of information helps the system capture facial features even under rapid head movement.

For feature tracking, we employ a Kalman filter for each feature. The Kalman filter imposes a smooth constraint on the motion of each feature, thereby alleviating the problem of local feature jumping. Each feature's motion state at each time instant (frame) can be characterized by its position and velocity. Let (x_t, y_t) represent its pixel position and (u_t, v_t) its velocity at time t in x and y directions. The state vector at time t can therefore be represented as $S_t = (x_t y_t u_t v_t)^T$. The system can be modeled as

$$\mathbf{S}_{t+1} = \Phi \mathbf{S}_t + \mathbf{W}_t \,, \qquad (4)$$

where Φ is the transition matrix and $\mathbf{W}_t$ represents system perturbation.

We further assume that a feature detector based on a Gabor wavelet estimates the feature position $\mathbf{O}_t = (\hat{x}_t, \hat{y}_t)^T$ at time t. Therefore, the measurement model in the form needed by the Kalman filter is

$$\mathbf{O}_t = H\mathbf{S}_t + \mathbf{V}_t \,, \qquad (5)$$

where H is the measurement matrix and $\mathbf{V}_t$ represents measurement uncertainty. Given the state model in Eq. 4 and measurement model in Eq. 5, as well as some initial conditions, the state vector $\mathbf{S}_{t+1}$, along with its covariance matrix Σ_{t+1}, can be updated frame by frame. This process consists of two steps: state prediction and state updating. For state prediction, the position prediction of each feature $P^k_{t+1} = (x^k_{t+1}, y^k_{t+1})^T$ can be obtained based on the state model in Eq. 4. Meanwhile, we also get the error covariance matrix Σ_{t+1} to represent the uncertainty of the current prediction from Kalman filtering. The covariance matrix may be used to limit the search area.

The Kalman filtering estimates the predicted location of the object based on assumed motion models and a smoothing motion constraint. Capturing voluntary or involuntary rapid head movement is a difficult task. This problem is resolved by combining the Kalman prediction with head motion prediction based on pupil motion. By linearly combining the head motion with the Kalman filtering, we can obtain a relatively accurate prediction of feature location even under rapid head movement. The final predicted position for each facial feature is

$$\widehat{P}_{t+1} = P^f_{t+1} + \begin{pmatrix} e^{-\sigma_{xx}} & 0 \\ 0 & e^{-\sigma_{yy}} \end{pmatrix} (P^k_{t+1} - P^f_{t+1}) \,, \qquad (6)$$

where the entire head motion $P^f_{t+1} = (x^f_{t+1}, y^f_{t+1})^T$ is the average of two pupil motions between consecutive frames. $P^k_{t+1} = (x^k_{t+1}, y^k_{t+1})^T$ is the predicted position of each feature from Kalman filtering. σ_{xx} and σ_{yy} are the first and second diagonal entries of the covariance matrix Σ_{t+1}, respectively. σ_{xx} and σ_{yy} represent the accuracy of $P^k_{t+1} = (x^k_{t+1}, y^k_{t+1})^T$. They serve as the weights to weigh the contribution of P^f_{t+1} and P^k_{t+1} to the combined prediction. Considering two extreme cases, zero and infinite, we use the exponential of $-\sigma$. When it is zero, which means no uncertainty for the predicted position, the predicted position gets P^k. When it is infinite and the predicted position is very uncertain, we use the position from the head motion.

3.2.2 Feature detection

The combined prediction provides a possible region (as determined by the covariance matrix) centered at the predicted position. The next step is to detect the feature point near the predicted position. The traditional approach is to search each pixel within the area to detect the optimal position. In tracking a large number of feature points, however, this process is time consuming and is not acceptable for real-time implementation. Here a fast detection method based on the phase-shift theory proposed by [61] is employed instead.

For the pixel $(\overrightarrow{\mathbf{x}})$ in the vicinity of the predicted position $(\overrightarrow{\mathbf{x}}')$, the phase shift of the Gabor coefficients $\Omega(\overrightarrow{\mathbf{x}})$ from $\overrightarrow{\mathbf{x}}'$ can approximately be compensated by the terms $\overrightarrow{d} \times \overrightarrow{k_n}$. The $\overrightarrow{d}$ indicates the displacement from the predicted position $(\overrightarrow{\mathbf{x}}')$. So the phase-sensitive similarity function of these two pixels can be

$$\mathbf{S} = \frac{\sum_n m_n m'_n cos(\phi_n - \phi'_n - \overrightarrow{d} \times \overrightarrow{k_n})}{\sqrt{\sum_n m_n^2 \sum_n m'^2_n}} \,, \qquad (7)$$

where m_n and ϕ_n indicate the amplitude and phase in the complex Gabor coefficients $\Gamma(\overrightarrow{\mathbf{x}})$, respectively.

The similarity function can be approximated by the Taylor expansion of the cosine term and ignoring orders greater than 2:

$$\mathbf{S} \approx \frac{\sum_n m_n m'_n [1 - 0.5(\phi_n - \phi'_n - \overrightarrow{d} \times \overrightarrow{k_n})^2]}{\sqrt{\sum_n m_n^2 \sum_n m'^2_n}} \,. \qquad (8)$$

By maximizing the above function, we can obtain the optimal displacement vector $\overrightarrow{d}_{opt}$ of the feature position:

$$\overrightarrow{d}_{opt} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \theta_x \\ \theta_y \end{pmatrix} \qquad (9)$$

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

$$\theta_x = \Sigma_n m_n m'_n k_{nx}(\phi_n - \phi'_n)\,,$$

$$\theta_y = \Sigma_n m_n m'_n k_{ny}(\phi_n - \phi'_n)\,,$$

$$\Gamma_{xx} = \Sigma_n m_n m'_n k_{nx} k_{nx}\,,$$

$$\Gamma_{xy} = \Sigma_n m_n m'_n k_{nx} k_{ny}\,,$$

$$\Gamma_{yx} = \Sigma_n m_n m'_n k_{nx} k_{ny}\,,$$

$$\Gamma_{yy} = \Sigma_n m_n m'_n k_{ny} k_{ny}\,.$$

Basically, the phase-sensitive similarity function can only determine the displacements up to a half-wavelength of the highest frequency kernel, which would be ±2 pixel area centered at the predicted position for $k = \pi/2$. But this range can be increased using a low-frequency kernel. Currently, a three-level coarse-to-fine approach is used, which can determine up to ±8 pixel displacements. For each feature only three displacement calculations are needed to determine the optimal position, which dramatically speeds up the detection process and makes real-time implementation possible. Our method is similar to the phase-based optical flow work by Fleet et al. [57].

4 Spatial-pattern extraction

4.1 Local graph warping

A facial expression consisting of fiducial points undergoes a variety of pattern changes. To capture the pattern changes, it is important to extract and verify spatial relationships among individually extracted features. This task involves two issues. One is nonrigid object tracking; the other is self-occlusion. So far, a great deal of research on facial feature tracking work has been conducted based on an assumption of either (1) head motion without facial expression change or (2) facial expression change with minimum out-of-plane head rotation. Most of the research has focused on one of the above two issues. But for real-world facial expressions, both issues should be taken into consideration. We must deal with intensity profile changes due to expression changes and self-occlusion due to excessive out-of-plane head motion. To tackle these issues, we propose a warping-based reliability propagation approach.

To accurately locate facial features under the two situations mentioned above, a facial feature should not be viewed as an isolated point. We divide the entire face into three local graphical objects: right eye graph, left eye graph, and lower face graph, as shown in Fig. 5. Each graphical object is characterized by its shape and attributes. The shape describes the topology and geometry of a graphical object; the attribute carries information about its different properties. Here the shape is represented by edges between features. The attribute consists of the intensity profile of the feature in the form of Gabor wavelet coefficients. The warping transformation was originally proposed by Beier [6] and is widely used in the filmmaking industry. It is a powerful tool for flexibly creating natural facial expressions. The warping of a graphical object is a 2D transformation that produces a continuous deformation from object $\mathbf{O}_1$ to object $\mathbf{O}_2$. The $\mathbf{O}_1$ is called the source object, and $\mathbf{O}_2$ is called the destination. Instead of the reverse mapping widely used in filmmaking, here we employ the forward mapping warping. We use spatial relations in the previous frame and extracted features in the current frame to verify other current features. Figure 6 illustrates one line-pair-based and uniform-scale warping in the form of forward mapping.

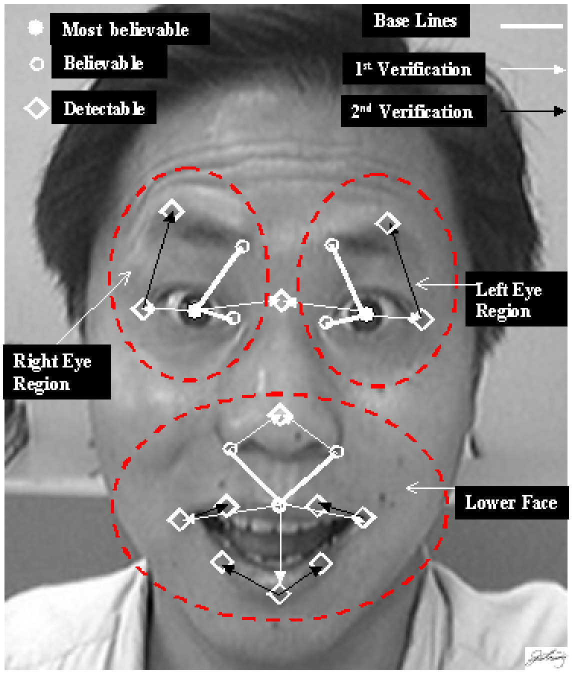

Fig. 5. Three local facial regions and classification of facial features for each region according to reliability

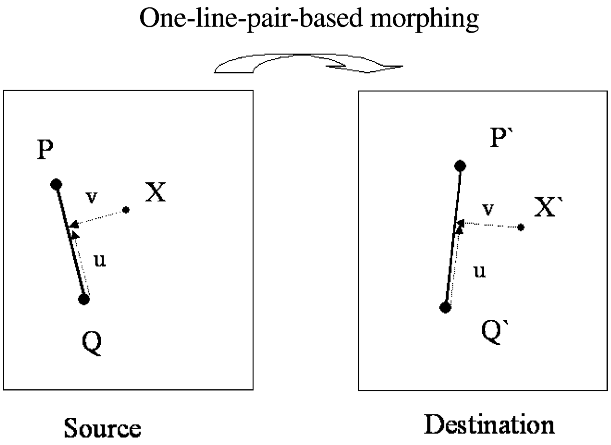

Fig. 6. One pair of line-segment-based warping

Given the interframe correspondence of the line pair $\overrightarrow{QP}$ and $\overrightarrow{Q'P'}$ and feature point X, the mapping position X' in the destination image from X is determined by the following formulae.

$$u = \frac{(\overrightarrow{XP}) \times (\overrightarrow{QP})}{\|\overrightarrow{QP}\|^2}\,, \tag{10}$$

$$v = \frac{(\overrightarrow{XP}) \times (\overrightarrow{QP})^{\perp}}{\|\overrightarrow{QP}\|^2}\,, \tag{11}$$

$$X' = P' + u(\overrightarrow{Q'P'}) + v(\overrightarrow{Q'P'})^{\perp}\,, \tag{12}$$

where $(\overrightarrow{QP})^{\perp}$ is a vector with the same length and perpendicular to $(\overrightarrow{QP})$.

With more than one pair of line segments, a weighting of the coordinate transformation for each line pair is performed. The weight based on the geometry such as distance to the line i and line length can be defined as

$$w_i = \left[\frac{\text{length}^p}{(a + \text{dist})} \right]^b , \tag{13}$$

where a, p, b are the factors to control the effects from the distance, length, and the line itself, respectively. The final mapping position of X is calculated as follows:

$$X' = \frac{\sum w_i X_i'}{\sum w_i} . \tag{14}$$

In the information-extraction step, the confidence level of the extracted feature is different. Besides the above geometric-based weighting, we add reliability-based weighting. We assign a larger b to the line segment with reliably extracted feature points. Based on the weighted warping, we can actively integrate the reliable features to verify and correct the unstable feature extraction. Warping's integrating ability makes it more flexible and powerful than similarity or affine transformation. Although the thin plate spline (TPS)-based transformation [3,4] is also an effective method for changing shapes, especially for biological subjects, the requirements for dense features and its iterated regularization make it inefficient for real-time applications [5]. Our feature-based warping needs only sparse features in each local graph and does not need time-consuming computation.

4.2 Refinement with reliability propagation

Each feature is related to one of three local graphs. With known pupil positions and a 22 landmark from the *common facial model*, the position of every other feature in the initial frame can be extracted and identified. From these extracted features their Gabor coefficients and spatial relationships are used to create a *personalized facialm* for the current face. During information extraction from image sequences, the Gabor coefficients as the template (or profile) of each feature are used to identify the feature point in the current frame. The Gabor coefficients are also updated frame by frame in order to handle change due to different expressions in the sequence. While tracking a nonrigid object, profile updating in each frame is a reasonable way to handle the profile change due to expression change and head motion. When occlusion of a specific feature happens, however, the updated profile becomes meaningless because in the current frame there is no detectable visual information for updating. This is a difficult issue. First, we have to update the profile frame by frame so as to capture the expression change. On the other hand, some nonsense profiles are adopted as templates to conduct tracking in the upcoming frame. The more critical issue is that basically it is difficult to detect when and where an occlusion will happen if we only focus on each single facial feature. Here a refinement method based on reliability propagation and warping is proposed. Although it is still hard to correctly locate the feature under the occlusion, we can detect when and where the occlusion will happen and immediately capture the position once the occluded feature shows up again.

All facial features are divided into three types: *most believable*, *believable*, and *detectable*, as shown in Fig. 5. The pupil positions detected from the IR-based sensor are the most reliable information. They labeled the *most believable* features, indicated by the filled circles. The inner feature points, such as inner corners of eyes and eyebrow, are not easily occluded. Features with high contrast such as nose ends are easily detected. These features – including the inner end of eyebrow, inner corner points of eye, nose ends, the upper-lip center of mouth – are labeled *believable* features, indicated by the unfilled circles. The remaining feature points, which are either vulnerable to the occlusion or difficult to detect stably, are labeled *detectable* features, indicated by the diamond shape. In order to conduct the warping-based process, two prime line segments connected by *most believable* or *believable* points are assigned within each local graph, denoted by the bold white lines in Fig. 5. The line segments that run from the upper-lip center of the mouth to each nose end form the prime lines in the lower face graph. The line segments running from the pupil to the inner corner of the eye, and from the pupil to the inner end of the eyebrow, form the prime lines in each eye graph.

Although each feature can perform independent movement during facial expression, we assume that all features belonging to the same local graph follow the similar interframe warping transformation within two consecutive frames. Since interframe correspondence of features and line segments has automatically been established by tracking, each *detectable* feature can be verified by the line-based warping. Specifically, the entire verification is conducted graph by graph and through two levels, shown in Fig. 7. In the first level, the prime lines in each local graph are used to create warping positions for the *detectable* features, which can be stably detected, except for occlusion. These features include the outer corners of eyes, nasal roots, the lower-lip centers of the mouth, the outer corners of the mouth, and the tip of the nose. In the second level, the verified *detectable* features in the first level are integrated to make new line segments. A more comprehensive weighting warping is used to verify very unstable features, such as the outer corners of the eyebrow, the middle points of the lower lip, and the middle points of the upper lip. In this way, reliably extracted features impose spatial constraints on the recovery of other unstable features. The constraint strength is ordered according to the feature's confidence level. The spatial relationship of each local graph dominated by the reliable features is propagated to unstable positions.

For the verification process, the warping position and tracked position of each *Detectable* feature are used to detect the tracking failure and determine the optimal position according to the steps in Fig. 8. First, the displacement between the warping position and tracked position is calculated. If the displacement is very large, the tracking failure of the feature due to the occlusion is detected. The warping position is adopted as the current position and the position of occluded feature is recovered. If the displacement is small, the Gabor similarities of the feature in the previous frame with the warping position and tracked position in the current frame are calculated, respectively. If the intensity profile at the warping position is more similar to one of the previous features, the warping position is assigned as the current position and the tracking error is corrected. Otherwise, the tracked position is preserved.

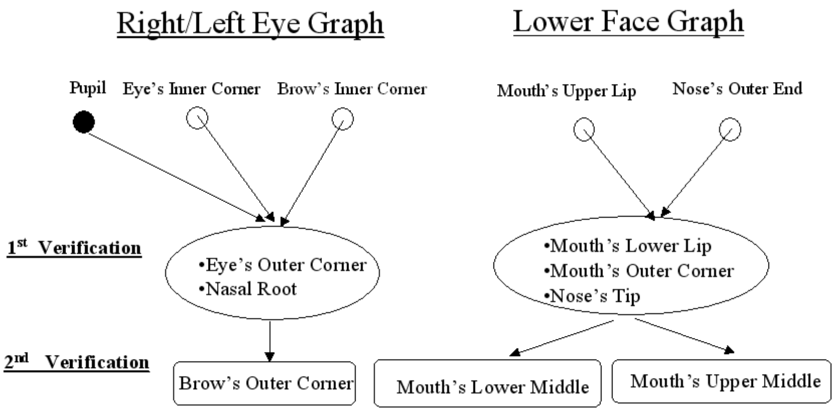

Fig. 7. Verification strategy with reliability propagation

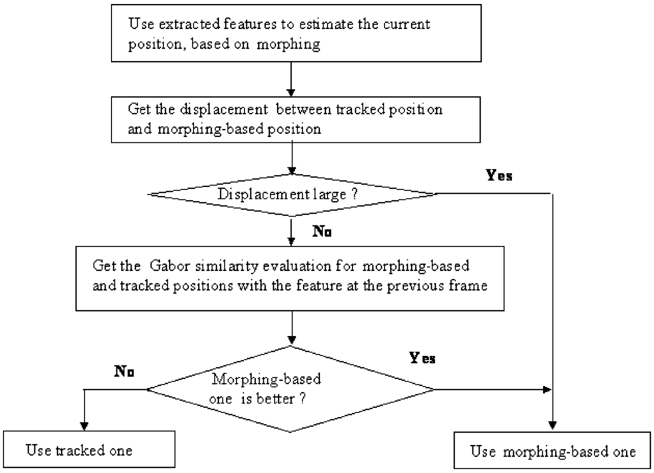
Use extracted features to estimate the current position, based on morphing
Get the displacement between tracked position and morphing-based position
Displacement large ?
Yes
No
Get the Gabor similarity evaluation for morphing-based and tracked positions with the feature at the previous frame
Morphing-based one is better ?
Yes
No
Use tracked one
Use morphing-based one

Fig. 8. Verification flowchart

5 Experimental results

5.1 Conversational sequence

Figure 9 shows a 600-frame sequence that displays the facial expression in a conversational scene. The facial expressions include smiling, yawning, talking, and blinking with out-of-plane head motions. The IR sensor provides the even and odd video sequences, corresponding to bright and dark pupil images, respectively. The odd sequence with detected pupil positions are used as input for the information-extraction process. Blinking is a constant facial event in all facial expression sequences. In this sequence, it happened 11 times. Each blinking period lasted about 4 to 6 frames. During these periods, the eyes closed. Except for the blinking periods, the IR-based pupil detection provided reliable pupil positions.

Table 1. The result of feature extraction

Pupil-Kalman approach	Total features	Failures features	Accuracy ratio
Without verification	13200	159	98.8%
With verification	13200	34	99.7%

Table 1 shows the result for extraction of all the facial features. We used the facial-feature position in the initial frame as the physical position of each facial landmark. The human operator checked each tracked feature in the upcoming frames with the physical position on the face to determine whether or not the feature extraction succeeded. In this sequence the head conducted smoothing motions and moderate out-of-plane head rotations. With the pupil constraints and Kalman filtering, most of the features were successfully extracted. The failures mainly occurred on the outer corners of the eye at the beginning or end of blinking. Because blinking is a kind of rapid local change, the intensity profile around the

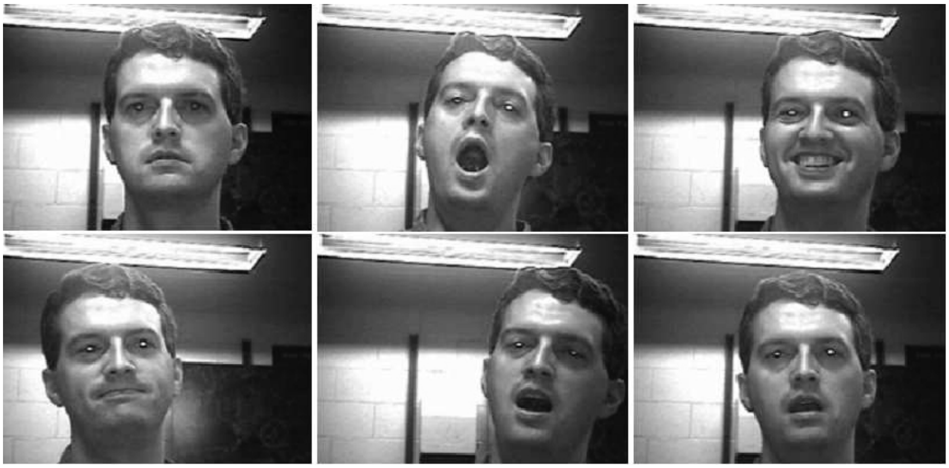

Fig. 9. Conversational sequence with bright-pupil effect

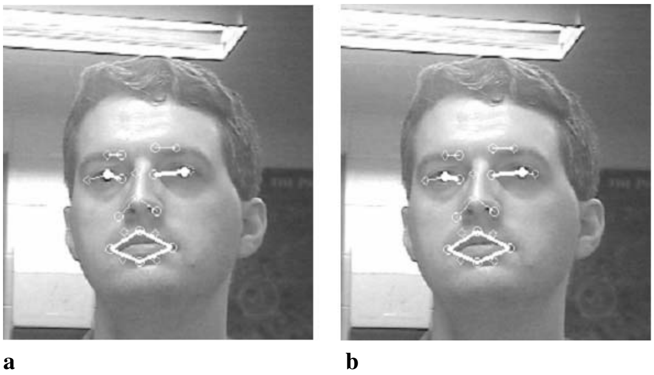

Fig. 10. Failure and modification for outer corner of left eye at frame 500

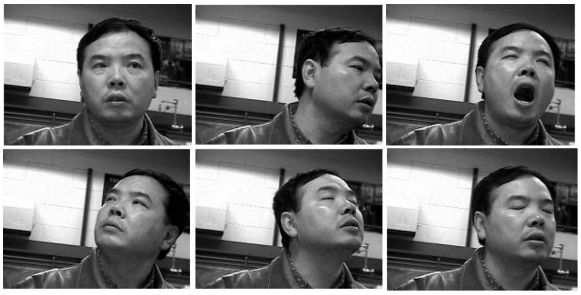

Fig. 11. A yawning facial expression sequence

outer corners of the eye undergoes rapid changes even within two consecutive frames.

Figure 10a shows the failure on the outer corner of the left eye (right side in the image plane). With the warping-based reliability propagation, the reliable neighbor features – pupil, inner corner of eye, and inner end of eyebrow – were used to verify the outer corner of the left eye. Since the position computed from the warping transformation was significantly different from the detected position, the result was corrected in Fig. 10b. The failure number of features was reduced by the reliability propagation, as shown in Table 1. The remaining failure is due mainly to the inconsistency between human judgment and the thresholds of verification in the reliability propagation. In the case of out-of-plane head motion, the human operator is often able to identify the physical position precisely only with great difficulty because the intensity profile of some features is significantly different from that in the initial frame.

5.2 Drowsy sequence

Figure 11 shows a typical sequence of a person's fatigue with significant facial expression changes and excessive out-of-plane head rotations. It consists of 449 frames. The person in the scene yawned from the neutral state, then moved the head rapidly from the frontal view to the large side view and back in the opposite direction, raised the head up, and finally returned to the neutral state. The head motion involved blended expressions.

Table 2. Comparison of three approaches

Approaches	Total features	Failure features	Accuracy ratio
Simple tracking	9856	854	91.3 %
Pupil-Kalman without verification	8960	174	98.0 %
Pupil-Kalman with verification	8960	64	99.3 %

Table 3. Failures in the simple tracking

Features	Failure periods [Start frame, End frame]	Lost features
LP	[54,234], [270,343]	253
LB(2)	[75,121]	92
MF(8)	[400,433]	264
OCR	[188,218]	30
RNE	[400,433]	33
OCL	[75,198],[378,433]	178
LNE	[399,403]	4

Table 4. Failures with constraints

Features	Failure periods [Start frame, End frame]	Lost features
LB(2)	[75,95]	40
OCL	[75,110], [373,448]	110
LNE	[398,422]	24

Table 5. Failures after refinement

Features	Failure periods [Start frame, End frame]	Lost features
LB(2)	[75,95]	40
LNE	[398,422]	24

Rows 3 and 4 in Table 2 report the results of our tracking method without and with reliability propagation, respectively. For comparison, we also show the result from a linear-motion-model-based tracking approach (called simple tracking in row 2). There are 22 facial features in each frame. In the simple tracking, the pupil looks like a normal tracking feature. So the total number of features is 9856. There were 854 features incorrectly extracted in the whole sequence. Table 3 shows the details of the failure.

The extraction failure happened on the left pupil (LP), left eyebrow (2 points) (LB), mouth features (8 points) (MF), outer corner of right eye (OCR), right nose end (RNE), outer corner of left eye (OCL), and left nose end (LNE). The main reasons for the failure are (1) unstable intensity information (LP, LB), (2) rapid head motion (MF, OCR, RNF), and (3) self-occlusion (OCL, LNF).

With the constraints from pupil positions and Kalman filtering, the number of extraction failures decreases significantly (Table 4). Since most pupil positions were detected by the IR active sensing, we remove the number of pupils from the total feature number. The total number is 8960. The extracting failure due to rapid head motions on MF, RNE, and OCR were improved.

With the warping-based reliability propagation further improvement has been reached (Table 5).

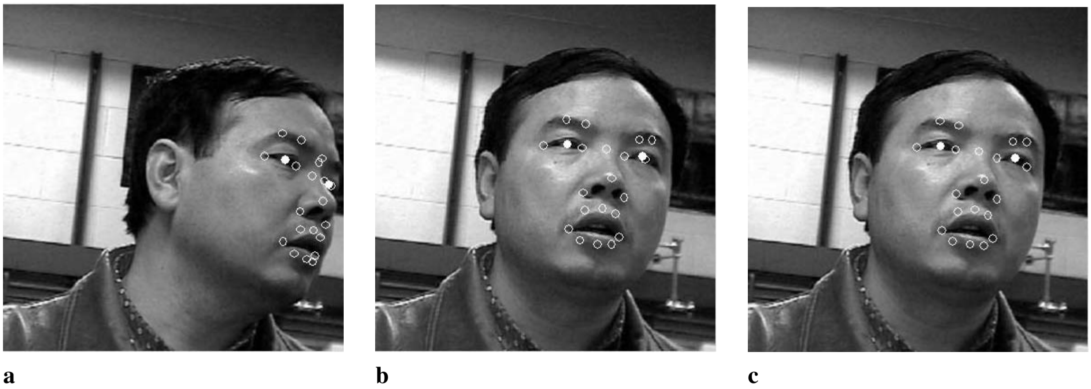

Fig. 12. a Occlusion on OCL at frame 70. **b** Result at frame 87 without verification. **c** Result at frame 87 with verification

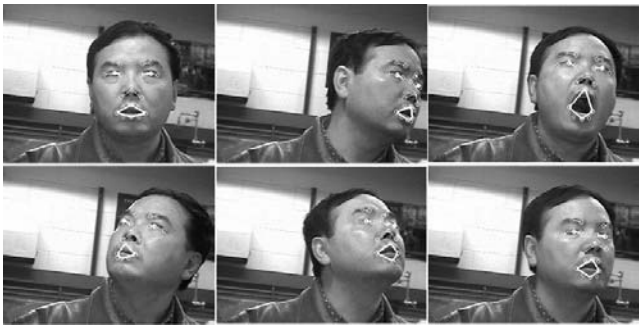

Fig. 13. Final result in form of local graphs

An example of self-occlusion is shown in Fig. 12. Figure 12a depicts the self-occlusion of the left eye at frame 70. Figure 12b, c show the tracked OCL at frame 87 without and with verification, respectively. The tracking failure due to the occlusion on the outer corner of left eye (OCL) was corrected by the warping-based verification.

Figure 13 displays the final extracted results in the form of local graphs.

6 Conclusion

In this paper, we proposed a robust approach to the information extraction of real-world facial expressions. The accurate results come from: (1) active sensing to help in the robust detection of pupils, (2) combination of the Kalman filtering with the pupil positions to effectively constrain feature locations, and (3) warping-based reliability propagation to handle occlusion and unstable features so as to robustly capture spatial relationships among features under excessive out-of-plane head rotations and significant expression changes. The extracted local graphs and their spatiotemporal relationships are used to conduct facial expression classification. Our facial-feature-tracking methods have been successfully applied to human fatigue monitoring [59], human emotion recognition [60], and animation [58]. Video demos of these applications may be found at `http://www.ecse.rpi.edu/homepages/cvrl/Demo/demo.html`.

Acknowledgements. This work is supported by a grant from the US Air Force Office of Scientific Research.

References

1. Abd-Almageed W, Fadali MS, Bebis G (2002) A non-intrusive Kalman filter-based tracker for pursuit eye movement. In: Proceedings of the 2002 American Control conference
2. Ahlberg J (2000) Real-time facial feature tracking using an active model with fast image warping. `http://citeseer.ist.psu.edu/538871.html`
3. Bookstein F (1989) Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell 11(6):567–585
4. Bookstein F (1991) Morphometric tools for landmark data. Oxford: Cambridge University Press, Cambridge, UK
5. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
6. Beier T, Neely S (1992) Feature-based image metamorphosiss. In: Computer Graphics (Proceedings of SIGGRAPH 92) 26(2):35–42
7. Bourel F, Chibelushi CC, Low AA (2000) Robust facial feature tracking. Proceedings of the 11th British machine vision conference, 1:232–241
8. Bretzner L, Lindeberg T (1998) Feature tracking with automatic selection of spatial scales. Comput Vis Image Understand 71(3):385–392
9. Chetverikov D, Verest'oy J (1998) Tracking feature points: a new algorithm. In: Proceedings of the international conference on pattern recognition, pp 1436–1438
10. Choi CS, Takebe T (1994) Analysis and synthesis of facial image sequences in model-based image coding. IEEE Trans Video Technol 4(6):257–275
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
12. Daugman J (1988) Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. IEEE Trans ASSP 36:1169–1179
13. Eisert P, Girod B (1998) Analyzing facial expressions for virtual conferencing IEEE Comput Graph Appl 18(5):70–78
14. Fieguth P (1997) Color-based tracking of heads and other mobile objects at video frame rates. In: Proceedings of the conference on computer vision and pattern recognition
15. Gorodnichy D (2002) On importance of nose for face tracking. In: Proceedings of the international conference on automatic face and gesture recognition (FG'2002)

16. Haro A, Flickner M, Essa I (2000) Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 163–168
17. Huang J, Ii D, Shao X, Wechsler H (1998) Pose discrimination and eye detection using support vector machines (svms), In: Proceedings of NATO-ASI on face recognition: from theory to applications, pp 528–536
18. Hwang V (1998) Tracking feature points in time-varying images using an opportunistic selection approach. Pattern Recog 22:247–256
19. Ji Q, Yang X (2001) Real time visual cues extraction for monitoring driver vigilance. In: Schiele B, Sagerer G (eds) Lecture notes in computer science, vol 2095. Springer, Berlin Heidelberg New York, p 107
20. Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 85:35–45
21. Kapoor A, Picard RW (2002) Real-time, fully automatic upper facial feature tracking. In: Proceedings of the 5th IEEE international conference on automatic face and gesture recognition, pp 10–16
22. Kass M, Witkin A, Terzopoulos D (1987) Snakes: active contour models. Int J Comput Vis 1:321–332
23. Lee T (1996) Image representation using 2d Gabor wavelets. IEEE Trans Pattern Anal Mach Intell 18(10):959–971
24. Li H, Roivainen P, Forchheimer R (1993) 3-D motion estimation in model-based facial image coding. IEEE Trans Pattern Anal Mach Intell 15(6):545–555
25. Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the international joint conference on artificial intelligence, pp 674–679
26. Luettin J, Thacker NA, Beet SW (1996) Locating and tracking facial speech features. In: Proceedings of the international conference on pattern recognition, pp 652–656
27. Malciu M, Preteux F (2000) Tracking facial features in video sequences using a deformable model-based approach. `http://citeseer.nj.nec.com/394625.html`
28. Manjunath B, Chellappa R, Malsburg C (1992) A feature based approach to face recognition. Proceedings of IEEE international conference on computer vision and pattern recognition, pp 373–378
29. Maurer T, Malsburg C (1996) Tracking and learning graphs on image sequences of faces. In: Proceedings of the international conference on artificial neural networks, pp 373–378
30. McKenna S, Gong S, Wurtz R, Tanner J, Bannin D (1997) Tracking facail feature points with gabor wavelets and shape models. In: Proceedings of the international conference on audio-and video-based biometric person authentication, pp 35–42
31. Meyer F, Bouthemy P (1994) Region-based tracking using affine motion models in long image sequences. CVGIP Image Understand 60:119–140
32. Morimoto C, Koons D, Amir A, Flickner M (1999) Framer-ate pupil detector and gaze tracker. In: Proceedings of IEEE ICCV, frame-rate workshop
33. Moriyama T, Kanade T, Cohn J, Xiao J, Ambadar Z, Gao J, Imanura M (2002) Automatic recognition of eye blinking in spontaneously occurring behavior. In: Proceedings of the international conference on pattern recognition (ICPR '2002)
34. Pantic M, Rothkrantz L (2000) Automatic analysis of facial expressions: The state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
35. Petajan E, Graf H (1996) Robust face feature analysis fo automatic speech-reading and character animation. In: Proceedings of the international conference on automatic face and gesture recognition, pp 357–362
36. Rangarajan K, Shah M (1991) Establishing motion correspondence. CVGIP: Image Understanding 54:56–73
37. Salari V, Sethi IK (1990) Feature point correspondence in the presence of occlusion. IEEE Trans Pattern Anal Mach Intell 12:87–91
38. Sethi IK, Jain R (1987) Finding trajectories of feature points in a monocular image sequence. IEEE Trans Pattern Anal Mach Intell 9:56–73
39. Shapiro L,Wang H, Brady J (1992) A matching and tracking strategy for independently-moving, non-rigid objects. In: Proceedings of the 3rd British machine vision conference, pp 306–315
40. Shi J, Tomasi C (1994) Good features to track. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 593–600
41. Sorenson HW (1970) Least-squares estimation: from Gauss to Kalman. IEEE Spectrum 7:63–68
42. Tomasi C, Kanade T (1991) Detection and tracking of point features. Technical Report, Carnegie Mellon University, Pittsburgh, PA
43. Thompson W, Lechleider P, Stuck E (1993) Detecting moving objects using the rigidity constraint. IEEE Trans Pattern Anal Mach Intell 15:162–166
44. Torresani L, Bregler C (2002) Space-time tracking. In: Proceedings of ECCV, pp 801–812
45. Tomasi C, Kanade T (1992) Shape and motion from image streams: a factorization method. Carnegie Mellon CMU-CS-92-104, pp 1–36
46. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
47. Yang J, Stiefelhagen R, Meier U, Waibel A (1998) Real-time face and facial feature tracking and applications. In: Proceedings of the international conference on auditory-visual speech processing (AVSP'98)
48. Yuille A, Cohen DS, Hallinan PW (1989) Feature extraction from faces using deformable templates. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 104–109
49. Yuille A, Hallinan P (1993) Deformable templates. In: Blake A, Yuille A (eds) Active vision. MIT Press, Cambridge, MA, pp 21–38
50. Zhang L (1998) Automatic adaptation of a face model using action units for semantic coding of videophone sequences. IEEE Trans Circuits Syst Video Technol 8(6):781–795
51. Zheng Q, Chellappa R (1995) Automatic feature point extraction and tracking in image sequences for arbitrary camera motion. Int J Comput Vis 15:31–76
52. Zhang Z (1998) Feature-based facial expression recognition: Experiments with a multi-layer perception. Technical report INRIA, no 3354
53. Zhong Y, Jain AK, Dubuisson-Jolly M (2000) Object tracking using deformable templates. IEEE Trans Pattern Anal Mach Intell 22(5):544–549
54. Zhu Z, Ji Q, Fujimura K, Lee K (2002) Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination. In: Proceedings of the international conference on pattern recognition, pp 373–378
55. Wang H, Brady J (1992) Corner detection with subpixel accuracy. Technical Report OUEL, Department of Engineering Science, University of Oxford, no 1925/92
56. Wiskott L, Fellous J, Krieger N, Malsburg C (1995) Face recognition and gender determination. In: Proceedings of the international workshop on automatic face and gesture recognition, pp 92–97

57. Fleet D, Jepson A (1993) Stability of phase information. IEEE Trans Pattern Anal Mach Intell 15(12):1253–1268
58. Wei X, Zhu Z, Yin L, Ji Q (2004) A real time face tracking and animation system In: 1st IEEE workshop on face processing in video, in conjunction with IEEE international conference on computer vision and pattern recognition
59. Gu H, Ji Q (2004) An automated face reader for fatigue detection, In: 6th international conference on automatic face and gesture recognition, 17–19 May 2004, Seoul, Korea
60. Li X, Ji Q (2003) Active affective state detection and assistance with dynamic Bayesian networks. 3rd workshop on affective and attitude user modeling assessing and adapting to user attitudes and affect: why, when and how?, in conjunction with 10th international conference on user modeling, Pittsburgh, PA
61. Wiskott L, Fellous J-M, Kruger N, von der Malsburg C (1997) Face recognition by elastic bunch graph matching. IEEE Trans Pattern Anal Mach Intell 19(7):775–779
62. Hutchinson TE (1990) Eye movement detection with improved calibration and speed. US patent 4,950,069
63. Cristinacce D, Cootes TF (2003) Facial feature detection using ADABOOST with shape constraints. In: Proceedings of BMVC, 1:231–240
64. Cristinacce D, Cootes TF (2004) A comparison of shape constrained facial feature detectors. In: Proceedings of the international conference on face and gesture recognition, pp 375–380
65. Cristinacce D, Cootes TF, Scott I (2004) A multistage approach to facial feature detection. In: Proceedings of the British machine vision conference, 1:277–286
66. Cootes TF, Edwards GJ, Taylor CJ (1998) Active appearance models. In: Proceedings of the European conference on computer vision 2:484–498

Haisong Gu received his Ph.D. in computer vision from Osaka University, Japan in 1993. After working as an assistant professor in Osaka University, he joined Matsushita Electric Works (Panasonic), Japan. Since 2002 he has been working at the University of Nevada, Reno. His research interests include computer vision, machine learning, image processing, and intelligent robot systems. He is a member of the IEEE.

Qiang Ji received his Ph.D. in electrical engineering from the University of Washington in 1998. He is currently an associate professor in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute in Troy, NY. His areas of research include computer vision, probabilistic reasoning for decision making and information fusion, pattern recognition, and robotics. Dr. Ji has published more than 60 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. His latest research focuses on face detection and recognition, facial expression analysis, image segmentation, object tracking, user affect modeling and recognition, and active information fusion for decision making under uncertainty. Dr. Ji is a senior member of the IEEE.