Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences

Yongmian Zhang, *Member*, *IEEE*, and Qiang Ji, *Senior Member*, *IEEE*

Abstract—This paper explores the use of multisensory information fusion technique with Dynamic Bayesian networks (DBNs) for modeling and understanding the temporal behaviors of facial expressions in image sequences. Our facial feature detection and tracking based on active IR illumination provides reliable visual information under variable lighting and head motion. Our approach to facial expression recognition lies in the proposed dynamic and probabilistic framework based on combining DBNs with Ekman's Facial Action Coding System (FACS) for systematically modeling the dynamic and stochastic behaviors of spontaneous facial expressions. The framework not only provides a coherent and unified hierarchical probabilistic framework to represent spatial and temporal information related to facial expressions, but also allows us to actively select the most informative visual cues from the available information sources to minimize the ambiguity in recognition. The recognition of facial expressions is accomplished by fusing not only from the current visual observations, but also from the previous visual evidences. Consequently, the recognition becomes more robust and accurate through explicitly modeling temporal behavior of facial expression. In this paper, we present the theoretical foundation underlying the proposed probabilistic and dynamic framework for facial expression modeling and understanding. Experimental results demonstrate that our approach can accurately and robustly recognize spontaneous facial expressions from an image sequence under different conditions.

Index Terms—Facial expression analysis, dynamic Bayesian networks, visual information fusion, active sensing.

1 Introduction

Facial expressions in spontaneous interactions are often characterized by the presences of 1) significant out-of-plane head motion, 2) temporal behaviors, and 3) partial occlusions. These aspects pose a number of technical challenges in developing facial expression recognition systems. Numerous techniques have been proposed for facial expression recognition within the past several years. However, much progress has been made toward recognizing face expressions that are mostly based on static face images without regard to partial occlusions. A facial expression is indeed the human behavior. It often reveals not only the nature of the deformation of facial features, but also the relative timing of facial actions as well as their temporal evolution. It is clearly of interest for human-computer interactions and human behavior analysis that an automated facial expression recognition system is capable of recognizing the facial actions, yet modeling their temporal behavior so that various stages of the development of a human emotion can be visually analyzed and dynamically interpreted by machine. More importantly, it is often the temporal changes that provide critical information about what we try to infer and understand about human emotions that possibly link to the facial expressions. Though efforts for facial expression recognition have been renewed most recently, the existing methods have typically focused on one part of the issues or the other, as discussed in the next section. Extending these systems to spontaneous facial behavior is a nontrivial problem of critical importance for realistic application of this technology [1].

The general goal of this research is to introduce a probabilistic and dynamic framework for spontaneous facial expression representation and recognition. The framework allows recognizing facial expression in spontaneous interactions from an image sequence. The following constitutes the framework based on which our approach is developed and it offers significant advantages over the techniques previously addressed in this field:

- *Facial Motion Measurement*: The measurement of facial motion is through tracking of facial features by simultaneously using an active Infra-Red (IR) illumination and Kalman Filtering. The pupil positions detected by using IR illumination are used to constrain the detection and tracking of other feature positions so that facial features can be robustly and accurately tracked under variable head motion and illumination condition.
- *Facial Expression Representation*: Facial expression representation integrates the dynamic Bayesian networks (DBNs) with the facial action units (AUs) from psychological views. The DBNs provide a coherent and unified hierarchical probabilistic framework to represent not only the probabilistic relations of facial expressions to the complex combination of facial AUs, but also temporal behaviors of facial expressions.
- *Facial Expression Recognition*: Facial expression recognition lies in a framework of dynamic and active multisensory information fusion. The framework

• *The authors are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, JEC 6003, 110 8th St., Troy, NY 12180. E-mail: zhang_y@cs.unr.edu, qji@ecse.rpi.edu.*

Manuscript received 22 Aug. 2003; revised 30 Aug. 2004; accepted 26 Oct. 2004; published online 11 Mar. 2005.
Recommended for acceptance by K. Daniilidis.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0243-0803.

facilitates dynamically modeling temporal evolution of facial expressions; therefore, it increases the robustness in handling occluded expressions and the uncertainties of feature measurements in dynamic imagery.

The main contributions of this work are as follows: First, we formulate the active and dynamic visual information fusion based on the extension of DBNs for online facial expression analysis in video sequences. Second, we cast the otherwise deterministic and static facial action code system (FACS) [2] in a dynamic and stochastic framework to account for image uncertainties and the dynamic nature of facial expressions. The most important property of this approach lies in the explicit modeling the dynamic and stochastic behaviors of facial actions and in the systematic and active integration sensory observations over time. Third, we propose various facial feature measures to relate to AUs.

This paper mainly emphasizes the following three issues: 1) systematically representing facial visual cues at different levels of abstraction by combining the DBNs with facial AUs from the psychology sources, 2) modeling temporal evolution and intensity of facial expressions to better understand the dynamic behavior of the human emotion, and 3) performing active and dynamic visual information fusion to increase robustness and efficiency in facial expression analysis. The remainder of this paper is organized as follows: In the next section, we present an overview of the existing works and identify the relation to ours. We give a brief introduction to our facial feature tracking in Section 3. Section 4 outlines the facial feature extraction and representation. Section 5 presents the modeling of facial expressions. Experimental results and analysis are presented in Section 6. The final section provides discussions and conclusions.

2 Previous Work

The interest in facial expression analysis for human-computer interactions has been around for a decade and substantial efforts were made in early 1990s [3], [4], [5], [6], [7], [8], [9], [10], [11]. However, a shortcoming of early works is their inability of performing automatic face and facial feature detection. In light of recent advances in image and video processing, various methods have been proposed for recognizing facial expression ever since. A thorough survey of the existing works can be found in [12], [13], [14].

2.1 Facial Feature Measurement

One of the fundamental issues about facial expression analysis is how to represent visual information so that the system can precisely reveal the subtle motion of facial activity. Most attempts on the representation of visual information for facial expression have focused on optical flow analysis from facial action [9], [15], [16], [17], where optical flow is used to either model muscle activities or estimate the displacements of feature points. Several facial expression recognition systems have employed the model-based techniques [18], [16], [19], where face images are mapped onto a physical model of the geometrical mesh by image warping. Recent years have seen the increasing use of feature geometrical analysis to represent visual facial information [20], [21], [22]. The facial movements are quantified by measuring the geometrical displacements of facial feature points between the current frame and the initial frame. In order to capture the information in the original image as much as possible to allow the classifier to discover the relevant features, a number of works applied holistic gray level analysis including principal components analysis [23], [24], Gabor wavelet analysis [25], Eigenface and Fisherface approach [26], template matching [27], etc.

Flow estimates are easily disturbed by the variation of lighting and nonrigid motion and are also sensitive to the inaccuracy of image registration and motion discontinuities. It is difficult to design a deterministic physical model that accurately represents facial geometrical properties and muscle activities. The holistic approach usually involves a time intensive training stage. The trained model is often unreliable for practical uses due to interpersonal variations, illumination variability, and also difficult to adapt to dynamic emotional sequences. The feature-based representation, on the other hand, requires accurate and reliable facial feature detection and tracking to accommodate the variation of illumination, significant head movement and rotation, as well as nonrigid feature change. We developed a real-time automatic facial feature-based tracking technique that can sufficiently fulfill these requirements [28].

2.2 Classification with Spatial Analysis

A significant amount of research on spatial analysis for facial expression recognition has focused on using Neural Networks (NNs) [24], [29], [30]. They differ mainly in their input facial data, which are either brightness distributions of feature regions, Gabor wavelet coefficients of feature points, principle components of facial images, or even an entire face image. More recently, Lyons et al. [25] presented a Gabor wavelet-based method. The facial feature points sampled from a sparse grid covering on the face are represented by a set of Gabor filters and are then combined to form a single feature vector. The principle components of the feature vectors from training images are further analyzed by linear discriminant analysis to form discriminant vectors. Finally, classification was performed by projecting the input vector of a test image along the discriminant vectors. Colmenarez et al. [31] presented a Bayesian probabilistic approach to recognizing the face and facial expression. Taking advantage of mutual benefit in similarity measure between face and facial expression, their method improves the recognition performance. Since each person in a database must have a trained model, this limits its realistic use in facial expression analysis. Pantic and Rothkrantz [21] used a dual-view face model (a frontal-view and a profile view) to extract facial features in order to reduce the ambiguities of face geometry. The extracted facial data is converted to a set of rule descriptors based on FACS. The classification of facial expressions is performed by comparing the AU-coded description of observed expression against the rule descriptors of six facial expressions.

The common limitation of the above works is that the recognition is performed by using static cues from still face images without considering the temporal behaviors of facial expressions. The psychological experiments by Bassili [32] have suggested that facial expressions are more accurately recognized from a dynamic image than from a single static image. The temporal information often reveals information about the underlying emotional states. For this purpose, our work concentrates on modeling the temporal behaviors of facial expressions from their dynamic appearances in an image sequence. Consequently, this sets our work apart from the above approaches.

2.3 Classification with Spatio-Temporal Analysis

There have been several attempts to track and recognize facial expressions over time. These works can be summarized from the following three major research groups. Yacoob and Davis [15] proposed a region tracking algorithm to integrate spatial and temporal information at each frame in an image sequence. The rigid and nonrigid facial motions are extracted by computing optical flow. Facial expression recognition follows the rules of basic actions of feature components and the rules of motion cues as described in [2], [32]. Rosenblum et al. [33] expanded the above work by using a radial basis function neural network structured as expression layer, facial feature layer, and motion direction sensitivity layer. The correlations between facial motion patterns and facial expressions are trained. Recognition is performed by integrating the information of past motion direction into the current response. Black and Yacoob [34] presented an approach with local parameterized flow models, where the affine model and planar model represent head motion and rotation, while the curvature model represents nonrigid motions of facial features on the eyebrows and mouth. A set of parameters estimated from the models are used to distinguish facial expressions. The above approaches required a facial expression having ideally three temporal segments: the beginning, apex, and ending. In spontaneous behavior, such segments are often hard to detect, especially when multiple expressions are involved in sequences.

Essa and Pentland [16] presented a system featuring both facial motion extraction and classification. The facial motion is estimated by using optical flow, which refines recursively based on the control theory. A physical face model is applied for modeling facial muscle actuation. Facial expression classification is based on the invariance between the motion energy template learned from ideal 2D motion views and the motion energy of the observed image. Since their approach needs a physical face model, the classification accuracy therefore relies on the validity of such a model. However, designing a deterministic physical model that can accurately reflect facial muscle activities appears difficult. Oliver et al. [35] applied a Hidden Markov model (HMM) on facial expression recognition based on the deformation of mouth shapes tracked in real-time. Each of the mouth-based expressions, e.g., sad and smile, is associated with an HMM trained by using the mouth feature vector. The facial expression is identified by computing the maximum likelihood of the input sequence with respect to all trained HMMs. However, only a part of the facial expressions have the characteristic pattern of mouth shape.

The works in [36] and [22] focus on recognizing facial AUs rather than facial expressions. Lien et al. [36] explored HMMs for facial AU recognition. Each AU or AU combination is assigned a specific HMM topology according to the pattern of feature motions. A directed link between states of the HMM represents the possible inherent transition from one facial state to another. An AU is identified if its associated HMM has the highest probability among all HMMs given a facial feature vector. Since each AU or AU combination associates with one HMM, the approach is infeasible for covering a great number of potential AU combinations involved in facial expressions. Tian et al. [22] presented an NN-based approach, in which two separate NNs are constructed for the upper face AUs and the lower face AUs. An AU combination within either the upper face AUs or the lower face AUs is treated as either a new upper face AU or a new lower face AU. The inputs to the NNs for both training and classification are the parametric descriptions of nontransient and transient facial features from their multistate face and facial component models. Unlike the AU recognition above, in facial expression analysis, an expression often constitutes the complex combinations among both the upper facial AUs and the lower facial AUs.

The current works on the spatio-temporal analysis for facial expression understanding, in our view, suffer the following shortcomings:

1. The facial motion information is obtained mostly by computing dense flow between successive image frames. As we previously remarked, dense flow computing itself suffers severe shortcomings.
2. The facial motion pattern has to be trained offline, whereas the trained model limits its reliability for realistic applications since facial expressions involve great interpersonal variations and a great number of possible facial AU combinations. In spontaneous behavior, the facial expressions are particularly difficult to be segmented by a neutral state in an observed image sequence.
3. Facial temporal information usually takes from three discrete expression states in an expression sequence: the beginning, the peak, and the end of the expression. The facial movement itself is not measured. Therefore, the existing approaches are not able to model the temporal evolution and the momentary intensity of an observed facial expression, which are indeed more informative in human behavior analysis.
4. The HMM can model uncertainties and time series, but it lacks the ability to represent induced and nontransitive dependencies. However, a facial expression consists of not only its temporal information, but also a great number of AU combinations and transient cues. Other methods, e.g., NNs, lack the sufficient expressive power to capture the dependencies, uncertainties, and temporal behaviors exhibited by facial expressions.
5. The appearance of wrinkles in certain regions of the face also provides crucial cues to deduce a facial expression. Besides the works on AU recognition in [36], [22], transient facial features have not been considered rigorously by the existing works on facial expression analysis.
6. Facial expressions often present occlusions and missing features caused by measurement errors, image noises, and head rotation, while these issues have not been addressed adequately by the existing works.

We explore the use of a multisensory information fusion technique with DBNs to overcome the above limitations. A summary of this work may be found in [37]. There are two major deviations from the previous works that merit being highlighted. First, we focus on modeling the temporal evolution and the momentary intensity of the expression that, in our belief, can lead to better understanding the dynamic behavior of human emotion. Second, we emphasize how to cast the culture and ethnic independent AU descriptions to a dynamic probabilistic model to account for both the uncertainty and the dynamics of the visual information and to alleviate the influence of interpersonal variations in facial expressions, which plagues many current facial expression understanding approaches.

Fig. 1. IR illumination-based camera system: (a) hardware setting, (b) an even field image with bright pupil, and (c) an odd field image with dark pupil.

3 Facial Feature Tracking

Numerous techniques have been proposed for facial feature tracking [38], [39], [40]. Our approach to facial feature tracking relies on an active IR illumination so that it can detect pupils under large variation in lighting and head orientation. The detected pupils are then used to constrain the possible positions of other facial features. Generally, our technique is more robust and accurate due to the use of active sensing. In addition, our technique is real time, fully automatic, without any manual involvement, and applicable to different people without any training. The existing techniques do not have all these benefits. Our technique on facial feature tracking, however, is based on the following assumptions: 1) the image has the face in mostly frontal view throughout the sequence, out-of-plane ($\pm$ 30 degrees) head rotation is allowed and 2) only a single face is present in the observed scene.

3.1 Pupil Detection and Tracking

Our approach to facial feature detection starts with pupil detection. To assist pupil detection, the person's face is illuminated with an IR illuminator, which produces the dark/bright pupil effect [41], [28]. Fig. 1a shows our active IR illumination-based camera system which consists of an IR sensitive camera and a number of IR LEDs evenly positioned on two concentric rings centered along the camera optical axis. The inner ring of LEDs is placed tightly close to the camera optical axis. When the LEDs in the inner ring are on, this configuration allows eyes to echo the IR light back exactly along its incoming path back into the camera, producing the bright pupil effect as shown in Fig. 1b. On the other hand, the outer ring of LEDs is relatively far apart from the camera, as shown in Fig. 1a, and, as a result, the light reflected by the eyes will not return to the camera, therefore producing the dark pupil effect as shown in Fig. 1c. A circuit is designed to synchronize the LED light sources on the inner ring and outer ring with the camera even and odd field scan so that the inner ring is on for the even field while the outer ring is on for the odd field. The interlaced image is subsequently deinterlaced by a video decoder to produce two images, namely the even field image and the odd field image, corresponding to the bright and dark pupil images, respectively, as shown in Fig. 1b and Fig. 1c. This property allows us to identify candidates of eye images from the difference image generated by subtracting odd image from the even image. Additional geometric and photometric properties (e.g., size, shape, and average intensity) are then used to eliminate spurious eye candidates. The detected eyes (pupils) are subsequently tracked by using a technique based on combining Kalman filtering [42] with mean shift [43].

3.2 Facial Feature Detection and Tracking

To characterize facial expressions, we propose to detect 26 facial features around the regions of eyes, nose, and mouth, as shown in Fig. 3. We perform facial feature detection and tracking on the odd field image since the odd field image still possesses normal gray levels. Given the detected eye positions, the initial regions of these features are located using some anthropometric statistics [44] that characterize the spatial relationships between eyes and nose and between nose and mouth corners, etc.

We represent the intensity distribution of each feature point x and its local neighborhood surrounding x with a set of multiscale and multiorientation Gabor wavelet coefficients. In our implementation, we have 18 Gabor wavelet coefficients represent each feature point along with an additional set of Gabor wavelet coefficients from its neighboring pixels.

The displacement of feature points in the face plane must be obtained through feature tracking. The image plane and the face plane do not have to always be the same. When the face is frontal, the two planes are the same. But when the face rotates (left or right), the two planes will be different. We can track features under head rotation. Since 3D face pose (pan, tilt, swing) can be estimated from the tracked features, we can use the pose information to correct the feature displacements to eliminate feature displacement distortions due to a foreshortening effect. The facial features change over time due to facial expressions and head motion, which involves a need of repeated estimation of the feature locations. The Kalman filter [42] is particularly well-suited to this application. The prediction based on Kalman filtering assumes smooth feature movement. However, the prediction will be off significantly if the face undergoes a sudden or rapid movement. To handle this issue, we propose to approximate the face movement with eye movement since eyes can be reliably detected in each frame. Specifically, the offset of head motion can be determined by utilizing pupil positions between two consecutive frames. Accordingly, the predicted feature position can be obtained from the detected pupil positions. The final predicted state vector should be based on properly combining the one from Kalman filtering with the one from eyes. Given the predicted feature position, we need to search the region near the prediction position to locate the actual feature position in the next image. This is the feature detection step, in which we use the phase-sensitive similarity by matching the Gabor kernel of two image points [45]. According to the theory of Kalman filtering, the final feature position is determined by combining the detected feature location with the predicted one.

Because of the simultaneous use of the Kalman Filtering and the information of detected pupil location, our facial tracking can provide a very good approximation for the future position of a feature point, even with significant head movements involved.

4 Facial Information Extraction

Our approach to facial expression understanding relies on the linguistic descriptions of facial expressions from psychological view. The unique relationships are established between the facial motion cues and facial expressions so that the facial motion can be directly adapted for automated facial expression understanding.

TABLE 1
A List of AUs that Are Related to Six Facial Expressions (Adapted from [2])

AU1  Inner Brow Raiser	AU2  Outer Brow Raiser	AU4  Brow Lowerer	AU5  Upper Lid Raiser	AU6  Cheek Raiser	AU7  Lid Tightener
AU9  Nose Wrinkler	AU10  Upper Lip Raiser	AU12  Lip Corner Puller	AU15  Lip Corner Depressor	AU16  Lower Lip Depressor	AU17  Chin Raiser
AU20  Lip Stretcher	AU23  Lip Tightener	AU24  Lip Pressor	AU25  Lips part	AU26  Jaw Drop	AU27  Mouth Stretch

4.1 Linguistic Descriptions of Facial Expressions

Facial expressions represent a visible consequence of facial muscle activity. Though there are numerous human emotions in our daily life, research in psychology has indicated that at least six basic expressions, including happiness, sadness, anger, disgust, fear, and surprise, are universally associated with distinct facial expressions [46]. It is generally believed that these expressions can be described linguistically using culture and ethnically independent AUs. Such AUs were developed by Ekman and Friesen in their FACS [2], where each AU is coded based on the facial muscle involvements. The FACS now becomes the leading standard for measuring facial expressions in the behavioral sciences and is widely accepted by researchers in the area of automated facial expression recognition [34], [21], [36], [22]. Likewise, we also adapt the AU-coded descriptions of facial expressions in the FACS to describe the six emotional expressions. To facilitate our introduction to facial expression modeling, we illustrate the facial AUs pertaining to the six expressions in Table 1, which is directly adapted from [2].

A facial expression is indeed the combination of AUs. We group AUs of facial expressions as primary AUs and auxiliary AUs. By the primary AUs, we mean those AUs or AU combinations that can be clearly classified as or are strongly pertinent to one of the six expressions without ambiguities. In contrast, an auxiliary AU is the one that can be only additively combined with primary AUs to provide supplementary support to the facial expression classification. Consequently, a facial expression contains primary AUs and auxiliary AUs. For example, AU9 (Nose Wrinkler) can be directly associated with an expression of disgust, while it is ambiguous to associate a single AU17 (Chin Raiser) with a disgust expression. When AU9 and AU17 appear simultaneously, the classification of this AU combination to a disgust expression then becomes more certain. Hence, AU9 is a primary AU of a disgust, while AU17 is an auxiliary AU of disgust. Table 2 gives a summary of primary AUs and auxiliary AUs associated with six expressions, which is our extension to Ekman's work in [2].

Naturally, combining primary AUs belonging to the same category increases the degree of belief in classifying to that category, as shown in Fig. 2a. However, combining primary AUs across different categories may result in: 1) a primary AU combination belonging to a different facial expression, e.g., the combination of AU1 (Inner Brow Raiser), a primary AU for sadness, and AU5 (Upper Lid Raiser), a primary AU for surprise, generates a primary AU combination for fear as illustrated in Fig. 2b and 2) increasing ambiguity, e.g., when AU26 (Jaw Drop), a primary AU for surprise, combines with AU1, a primary AU for sadness, the degree of belief in surprise is reduced and the ambiguity of classification may be increased as shown in Fig. 2c. These relations and uncertainties are systematically represented by a probabilistic framework presented in Section 5. In principle, our approach allows a facial expression to be a probabilistic combination of any relevant facial AUs.

Additionally, the movement of facial transient features, such as the changes of wrinkles and furrows, also provides support cues to infer certain expressions. For example, a smiling face may lengthen and deepen the horizontal wrinkles on the eye outer canthi, while the vertical furrows between the eye brows tend to be intensive when a person expresses strong anger. The appearance of facial transient features is influenced by not only the interpersonal variation, but also by the age as well. Some of these transient features may become permanent due to age. We only consider the changes of these features as support evidence, which may partially contribute to the identification of facial expressions. Table 2 contains the transient features associated with facial expressions.

4.2 Facial Feature Extraction

The AUs of FACS described by the activation of the muscles is for human observers. In order to automatically extract the activation of the muscles from a face image, we have to quantitatively code AUs into facial feature movements that can be extracted directly from the face image. Fig. 3 presents facial geometrical relationships and furrow regions, where the feature points are located by face tracking throughout the image sequence.

Automated measuring of the geometrical displacement of facial features is analogous to human observations of facial activities. In order to adapt the FACS description for machine recognition of facial expressions, there is a need to establish a unique association between changes of the feature points and the corresponding AUs. At present, the association of the AUs and the corresponding movements

TABLE 2
The Association of Six Emotional Expressions to AUs, AU Combinations, and Transient Features

Emotional	Primary Visual Cues					Auxiliary Visual Cues					
Category	AU	AU	AU	AU	AU	AU	AU	AU	AU	AU	Transient Feature(s)
Happiness	6	12				25	26	16			wrinkles on outer eye canthi presence of nasolabial furrow
Sadness	1	15	17			4	7	25	26		
Disgust	9	10				17	25	26			presence of nasolabial furrow
Surprise	5	26	27	1+2							furrows on the forehead
Anger	2	4	7	23	24	17	25	26	16		vertical furrows between brows
Fear	20	1+5	5+7			4	5	7	25	26	

of the feature points are determined manually. Techniques (e.g., [22]) have been suggested to automatically build the association. The current automated techniques, however, are lacking robustness and accuracy. Correct association is crucial for accurate facial expression interpretation. Therefore, we manually establish the association of the AUs and the movements of the facial feature points as shown in Table 3 so that the facial visual changes are automatically measurable on imagery. Since this manual association can be done offline, it will not affect online performance.

The positions of the eye inner canthi (F and F' in Fig. 3) are the most stable to the facial actions among all feature points. We therefore choose these two fiducial points as a reference to measure the displacement of other feature points in Table 3 so that the accuracy of feature measurement can be ensured. Additionally, the symmetric property of the human face allows us to generate the redundant visual information for some feature points (e.g., feature points surrounding the eyes) which can be used to reduce the information uncertainty possibly resulted from missing features, inaccurate tracking, or partial occlusions. Some feature points, such as the eye outer canthi, are vulnerable to be occluded by head rotation and, thus, we use more than one way to measure the feature displacements related to these points. Failure of one way is therefore compensated by another way. Take AU4 (Brow Lower) for example. We measure not only $\angle HFI$ (see Fig. 3), but also the movement about the vertex of the upper eyelids as illustrated in Fig. 4.

The activation of facial muscles also produces transient wrinkles and furrows perpendicular to the muscular motion direction in certain face regions. For example, raising the outer brows wrinkles up one's frontal eminence and raising the cheeks may deepen the nasolabial fold and deform its initial shape. While one's forehead, nasolabial region, and eye corners may be furrowed with age and become permanent facial features, more or less, facial muscle movement causes changes in their appearance, such as deepening or lengthening. The transient features can therefore provide additional visual cues to support the recognition of facial expressions. The regions of facial wrinkles and furrows, as shown in Fig. 3, are located by the facial feature positions through feature tracking. The change of wrinkles in the region $\Box X$ is directly related to AU9 (Nose Wrinkler), while others merely enhance the identification of AUs, e.g., furrows in the regions $\Box Z$, $\Box Y$, $\Box V$, $\Box U$ provide diagnostic information for the identification of AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), and AU17 (Chin Raiser), respectively. The transient feature regions are summarized in Table 4. The presence of furrows and wrinkles on an observed face image can be determined by edge feature analysis in the areas that transient features appear. Similarly to [22], we use a Canny edge operator to quantify the intensity of furrows. Fig. 5 shows examples of transient feature detection. Besides the furrows in the nasolabial region, the change of the transient feature is quantified by the ratio of the number of edge pixels of the current image frame to that in its neutral state. The furrow is present if the ratio is over a predefined threshold. If there are a few hairs on the forehead, this technique can still work since hairs are mostly represented by vertical edges while forehead wrinkles are mostly horizontal edges.

We achieve the detection of nasolabial folds in two steps. First, edge points are extracted by using a Canny operator in a nasolabial region. Starting from tracked feature point P or P' (see Fig. 3), which is a junction of the nose and the nasolabial fold on the upper left corner of the nasolabial region as shown in Fig. 6, we search for the longest edge in

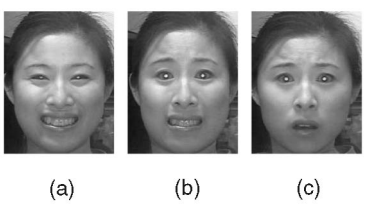


Fig. 2. Examples of AU combination: (a) AU12+AU6 (two primary AUs from the same category) enhances classification to happiness, (b) AU1+AU5 (two primary AUs from different categories) becomes a fear, and (c) AU26+AU1 (two primary AUs from different categories) increases ambiguity between a surprise and a fear.

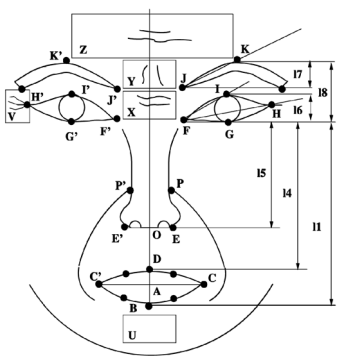


Fig. 3. The geometrical relationships of facial feature points, where the rectangles represent the regions of furrows and wrinkles.

TABLE 3
Motion-Based Feature Descriptions for AUs

AUs	Facial Visual Cues	Visual Channel(s)
AU1	$\angle FHJ$, $\overline{JF}$ increased OR	
	$\overline{JF}$ increased, $l8$ nonincreased	Brow
AU2	$l8$ increased and $\overline{JF}$ nonincreased	
	furrow in $\square Z$ increased	Brow, Wrinkler
AU4	$l8$, $\overline{FJ}$, $\overline{JJ'}$, $\overline{FP}$, $\overline{F'P'}$ decreased,	
	$\angle HFI$ increased and	
	wrinkle in $\square Y$ increased	Brow, Wrinkler
AU5	$l6$, $\overline{JF}$ and $\overline{JJ'}$ increased	Lid
AU6	nasolabial furrow presence and	Nasolabial,
	wrinkle in $\square V$ increased	Wrinkler
AU7	$\angle HFI$ nonincreased and	
	$\angle HGF$ increased	Lid
AU9	wrinkle increased in $\square X$	
	nasolabial furrow presence OR	Wrinkler,
	$\overline{PF}$, $\overline{FJ}$ decreased	Nasolabial
AU10	$l4$ decreased and $\lvert\overline{FC} - \overline{F'C'}\rvert$	
	increased, nasolabial presence OR	Lip,
	$\overline{OD}$ decreased, $\overline{DB}$, $\overline{C'C}$ increased	Nasolabial
AU12	$\overline{FC}$, $\overline{F'C'}$ decreased,	
	$\overline{CC'}$ increased, $\overline{GI}$ nonincreased	LipCorner
AU15	$\overline{FC}$, $\overline{F'C'}$, $\overline{CC'}$ increased	LipCorner
AU16	$\overline{OD}$ non-change, $\overline{DB}$ decreased	Lip
AU17	$\overline{OB}$ decreased and	
	wrinkle in $\square U$ presence	Chin, Wrinkler
AU20	$\overline{CC'}$ increased and	
	$\overline{FC}$, $\overline{F'C'}$ nonchange	LipCorner
AU23	$\overline{DB}$, $\overline{CC'}$ decreased	Lip
AU24	$\overline{DB}$ decreased, $\overline{CC'}$ nonchange	Lip
AU25	$\overline{DB}$ increased, $\overline{DB} < T_1$,	
	$\overline{CC'}$ nonincreased	Mouth
AU26	$T_1 < \overline{DB} < T_2$, $\overline{CC'}$ nonincreased	Mouth
AU27	$\overline{DB} > T_2$, $\overline{CC'}$ nonincreased	Mouth

Note: T_1 and T_2 are predefined thresholds of the mouth vertical span, which are determined separately for each person in the neutral state.

the nasolabial region as a nasolabial fold. Since the nasolabial fold has a generally preknown direction, we first search an edge point in the direction. If the edge point is not found in the known direction, we then try the neighbors of that edge point. The nasolabial fold is detected if the number of connected edge pixels is larger than a predefined threshold. However, the nasolabial wrinkle detection does not work for a bearded face.

(a) (b)

Fig. 4. The vertex of the upper eyelid usually shifts toward the inner eye canthus as a subject performs an anger: (a) neutral state and (b) an anger.

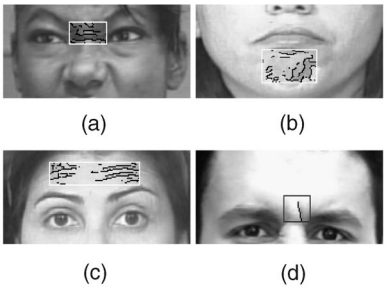
TABLE 4
Descriptions of Facial Transient Feature Regions

Transient Feature	Region in Fig.3	Feature Region Descriptions	Related to AU(s)
Forehead Wrinkle	$\square Z$	Rectangle: $\overline{KK'} \times 1/3\overline{KK'}$; k' and k are tracked	AU2
Nasal Wrinkle	$\square X$	Rectangle: $\overline{FF'} \times min(\overline{F'J'}, \overline{FJ})$; F, F', J, J' are tracked	AU9
Brow Wrinkle	$\square Y$	Square: $\overline{JJ'} \times \overline{JJ'}$; J, J' are tracked	AU4
Eye Wrinkle	$\square V$	Square: $\overline{IG'} \times \overline{IG'}$; I, G, H, I', G', H' are tracked	AU6
Chin Wrinkle	$\square U$	Rectangle: $\overline{CC'} \times 2/3\overline{CC'}$; C, C', A are tracked	AU17
Nasolabial Fold		refer to Fig. 6 and Fig. 7	AU6, AU9 AU10

The contraction or the extension of facial muscles may deform the initial nasolabial fold to a particular shape, as depicted in Fig. 6. We approximate the shape of a nasolabial fold as a quadratic of the form $y = ax^2 + bx + c$, where the coefficients of y can be obtained by fitting a set of the detected edge points to y in a least-squares sense. The coefficient a measures the curvature of nasolabial fold. There are two visual cues exhibited by a nasolabial fold that may explicitly cause cheek raise and, consequently, produce the AU6 (Cheek Raiser): 1) $a > 0$, as shown by $\widehat{PF}$ in Fig. 6, and 2) $a < 0$ and the nasolabial fold has a vertex, as shown by $\widehat{PE}$ in Fig. 6, that is, $x = -b/2a \in E$, where E is the set of edge points of the nasolabial fold. On the other hand, a nasolabial fold is a support evidence to AU9 (Nose Wrinkler) and AU10 (Upper Lip Raiser) if it has no vertex and $a < 0$, as shown by $\widehat{PD}$ in Fig. 6. The examples of nasolabial fold detection and modeling are presented in Fig. 7, where the coefficient a is -0.030116, 0.040323, and -0.027236, respectively.

5 Modeling Facial Expressions

In this section, we create a BN model to represent the causal relations between the facial expressions and facial AUs based on their linguistic descriptions as previously described. We then extend the BN to a DBN model for modeling the dynamic behaviors of facial expressions in



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

Fig. 5. Transient feature detection: (a) horizontal wrinkles between eyes, (b) furrows on the chin, (c) horizontal wrinkles on the forehead, and (d) vertical furrows between brows. Note that the original images are given in Table 1.

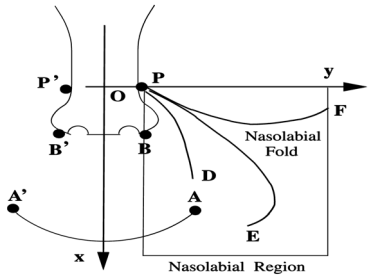


Fig. 6. Possible shapes of a nasolabial fold due to facial expressions.

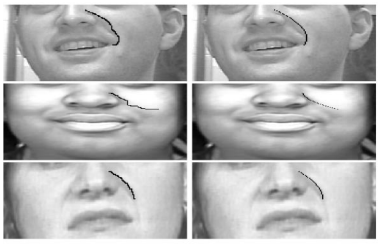


Fig. 7. The example results of nasolabial fold detection and modeling: (left column) detected nasolabial fold and (right column) fitting to a quadratic

image sequences. Finally, the theory of active sensing is theoretically formulated to efficiently and timely infer and recognize facial expressions in image sequences.

5.1 Modeling Facial Expressions with Bayesian Networks

The design of the BN causal structure should best reflect experts' understanding of the domain. According to the causal relations between AUs and the six expression categories in Table 2, we build the BN model as shown in Fig. 8, which best represents facial expressions for static face images. Our BN model of facial expression consists of three primary layers, namely, classification layer, facial AU layer, and sensory data layer.

The classification layer consists of a class (hypothesis) variable C including six states $c_1, c_2, \cdots, c_6$, which represent $happiness$, $sadness$, $disgust$, $surprise$, $anger$, and $fear$, respectively, and a set of attribute variables denoted as HAP, ANG, SAD, DIS, SUP, and FEA corresponding to the six facial expressions as shown in Fig. 8. The goal of this level of abstraction is to find the probability of class state c_i, which represents the chance of class state c_i given facial observations. When this probability is maximal, it has the largest chance that the observed facial expression belongs to the state of class variable c_i.

The AU layer is analogous to linguistic description of the relation between AUs and facial expressions in Table 2. Each expression category, which is actually an attribute node in the classification layer, consists of primary AUs and auxiliary AUs. A primary AU contributes stronger visual cues to the understanding of the facial expression than an auxiliary AU does. Hence, the likelihood of primary AUs to the facial expression is higher than that of auxiliary AUs.

The lowest level of layer in the model is the sensory data layer containing visual information variables, such as *Brows*, *Lips*, *Lip Corners*, *Eyelids*, *Cheeks*, *Chin*, *Mouth*, *Nasolabial Furrow*, and *Wrinkles*. All variables in this layer are observable. The visual observations are the facial feature measurements as summarized in Table 3 and Table 4. In our implementation, for facial features in pair such as brows, eyelids, eye wrinkle, and nasolabial fold, we measure the feature change on both sides of face and consider the one with the most prominent change as an evidence. Consequently, additional robustness can be achieved in handling partial occlusions or missing features.

The benefit of BN is that it allows us to include hidden layers between visual cues and the facial expressions. On the other hand, these hidden layers allow modeling correlations among visual cues. The direct mapping of the visual cues onto the expressions assumes conditional independence among visual cues, which is apparently not reasonable. Furthermore, it is also hard to specify the

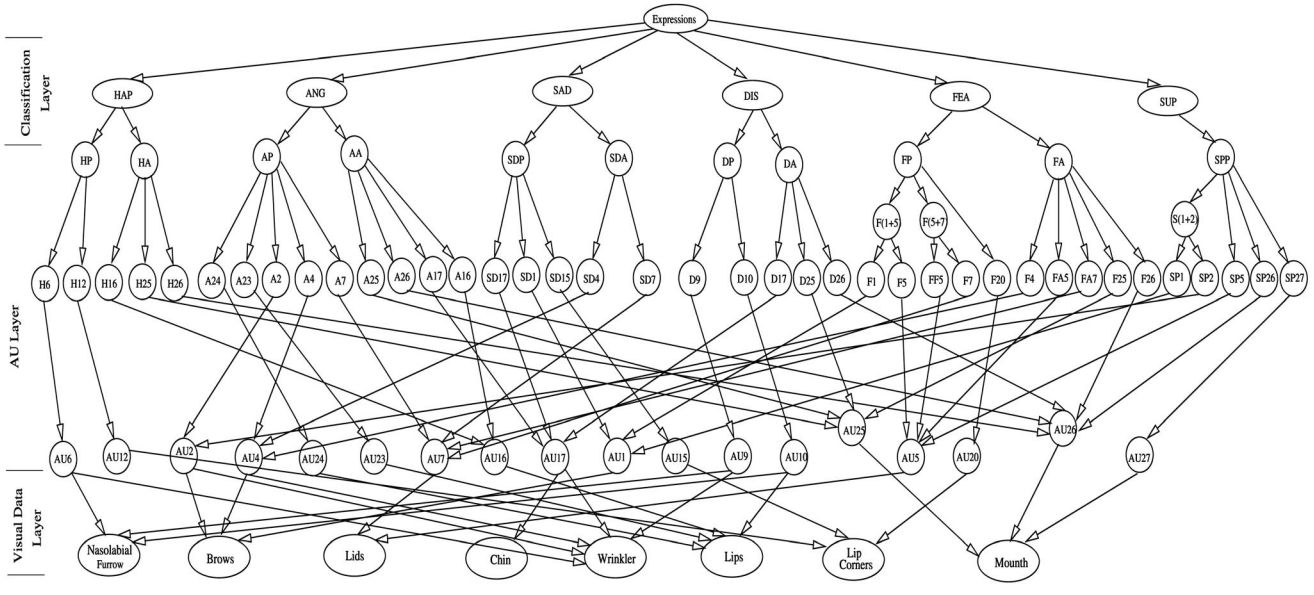


Fig. 8. The BN model of six basic emotional expressions. Note: HAP—Happiness. SAD—Sadness. ANG—Anger. SUP—Surprise. DIS—Disgust. FEA—Fear. The node HP denotes primary AUs for Happiness, while HA denotes auxiliary AUs for happiness. H6 means AU6 belonging to happiness. $F(1+5)$ denotes the combination of AU1 and AU5, which belongs to a fear. Other notations in the figure follow the same convention above.

conditional probabilities between the visual cues and their corresponding expressions. A typical example is that a visual cue AU5 (Upper Lid Raiser) may cause a fear, as shown in Fig. 8. However, AU5 can be further combined with other AUs, such as AU1 (Inner Brow Raiser), AU7 (Lid Tightener), and cause a fear in different degrees of belief. The direct mapping is difficult to reflect such details. Because of explicitly accounting for the correlations among visual cues, we can conclude that our structure of BN model has advantages compared with the direct mapping of the visual cues onto the expressions and should perform better.

5.2 Parameters of Facial Expression Model

Since the relationship between facial motion behaviors and the facial expressions is determined uniquely by human physiology and anatomy, theoretically, this alleviates the influence of interpersonal variation on facial expression model. Hence, the topology of the BN facial expression model is invariant over time. Nevertheless, the model needs to be parameterized by the necessary prior probabilities for the root nodes and the conditional probabilities for the intermediate nodes. The conditional probabilities of primary AUs or AU combinations for a given facial expression are based on the statistic results produced by a group of AU coders through visual inspection of AUs and their combinations, as presented in Table 2. Some of such data produced by the certified FACS coders have been reported in [21]. Since a single auxiliary AU lacks the sufficient visual information that can be used to relate it to a specific facial expression, it is difficult to statistically obtain its probability distribution for a facial expression. Take AU9 (Nose Wrinkler) and AU17 (Chin Raiser) for example. The probability of AU9+AU17 and AU9 given the expression of disgust can be assessed by a group of AU coders, which are 92 percent and 88 percent, respectively. However, the probability of AU17 given disgust can hardly be judged "by eye." We use the following formulae to estimate the conditional probabilities of auxiliary AUs given a facial expression. Let X_1 be a primary AU and X_2 be an auxiliary AU, and let $X = x$ be a facial expression which is a combination of X_1 and X_2. The probability of $X = x$ given an AU combination of X_1 and X_2 can be expressed under the assumption of independence among AUs

$$\begin{aligned} p(X = x \mid X_1, X_2) &= \frac{p(X = x, X_1, X_2)}{\sum_X p(X = x, X_1, X_2)} \\ &= \frac{p(X_1 \mid x)p(X_2 \mid x)}{p(X_1 \mid x)p(X_2 \mid x) + \frac{p(\neg x)}{p(x)} p(X_1 \mid \neg x)p(X_2 \mid \neg x)}. \end{aligned} \tag{1}$$

For notational convenience, let $q_1 = p(X_1 \mid X = x)$ be the probability of the primary AU X_1 given $X = x$, $q_2 = p(X_2 \mid X = x)$ be the probability of an auxiliary AU X_2 given $X = x$, and $q_{1,2} = p(X = x \mid X_1, X_2)$ as the probability of $X = x$ given X_1 and X_2. Assume that $p(X = x) = p(X = \neg x)$ (equal likely). Rewriting (1) yields,

$$q_2 = \frac{q_{1,2}(1 - q_1)}{q_1 + q_{1,2} + 2q_1 q_{1,2}}. \tag{2}$$

If $q_2 > 50$ percent, the auxiliary AU and the primary AU are in synergy enhancement. Otherwise, they are not. Again, take AU9 and AU17 for example. The conditional probability of AU9+AU17 given disgust is $q_{1,2} = 92$ percent, where AU9 and AU17 are a primary AU and an auxiliary AU of the expression of disgust, respectively. The conditional probability of AU9 given a disgust is $q_1 = 88$ percent. Hence, by (2), the conditional probability of AU17 given a disgust is $q_2 = 61.1$ percent, which indicates how much AU17 can possibly support AU9 to be the expression of a disgust when AU9 and AU17 appear simultaneously in a facial expression.

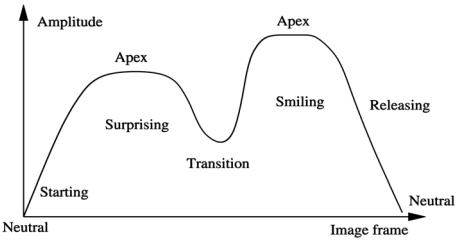


Fig. 9. An illustration that an expression sequence contains two emotional expressions, a surprise followed by a smile.

The parameters of conditional probabilities in the classification layer and sensory data layer (see Fig. 8) are estimated by Maximum Likelihood Estimation from the given sensory data. Since each conditional probability in the AU layer is known as described above, this guarantees to lead the rest of parameters converge to a local maximum of the likelihood surface.

5.3 Dynamic Modeling of Facial Expression

Facial expressions can be said to express emotions and emotions vary according to subject-environment interaction. As illustrated in Fig. 9, an expression sequence, in many cases, sequentially contains multiple expressions of different intensities due to evolution of the subject's emotion over time. The duration of facial expression is often related to the intensity of the emotion underlying the expression. Modeling such temporal behaviors of facial expressions allows better understanding of the human emotion at each stage of its development.

The static BN model of facial expression works with visual evidences and beliefs from a single time instant, and it lacks the ability to express temporal dependencies between the consecutive occurrences of expression in image sequences. To overcome this limitation, we use DBNs to model the dynamic aspect of a facial expression. The DBNs have previously been used successfully for hand gesture recognition [47]. Our DBN model is made up of interconnected time slices of static Bayesian networks (SBNs) described above, and the relationships between two neighboring time slices are linked by the first order Hidden Markov model [48], i.e., random variables at time t are affected by observable variables at time t, as well as by the corresponding random variables at time $t - 1$ only. The relative timing of facial actions during the emotional evolution is described by moving a time frame in accordance with the frame motion of a video sequence, so that the visual information at the previous time provides diagnostic support for current expression hypothesis. Eventually, the belief of the current hypothesis is inferred relying on the combined information of current visual cues through causal dependencies in the current time slice, as well as the preceding evidences through temporal dependencies. Fig. 10 shows the temporal dependencies by linking the top nodes of SBN in Fig. 8. Consequently, the expression hypothesis from the preceding time slice serves as a prior information for current

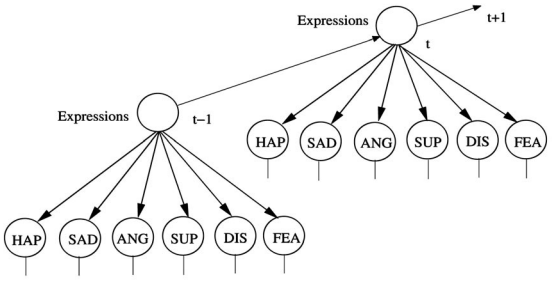


Fig. 10. The temporal links of DBN for modeling facial expression (two time slices are shown). Node notations are given in Fig. 8.

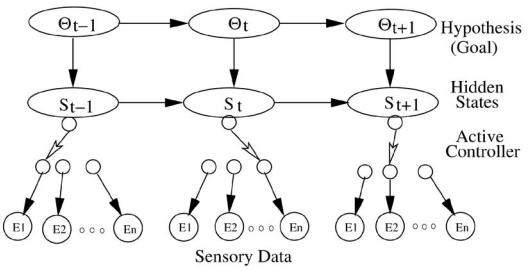


Fig. 11. A conceptual framework of DBN-based active information fusion. The system consists of a Goal, Hidden States, an Active Sensor Controller, and Sensory Data.

hypothesis, and the prior information is integrated with current data to produce a posterior estimate of current facial expression. In terms of Bayesian networks, the probability that we are interested in is the posterior distribution of current hypothesis of facial expressions Θ given a set of visual observations E, i.e., $p(\Theta \mid E)$. Applying Bayes' theorem and marginalization, we have

$$p(\Theta_t \mid E_t) = \frac{\sum_S p(\Theta_t, S_t, E_t)}{\sum_\Theta \sum_S p(\Theta_t, S_t, E_t)}, \tag{3}$$

where t is the discrete time indexm, S_t represents every configuration of hidden states, E_t represents current available facial visual information, and Θ_t is the current hypothesis of facial expressions. The Markov assumptions lead to the following expression for the joint distribution of states and visual observations by the probability chain rule

$$\begin{aligned} p(\Theta_t, S_t, E_t) &= p(\Theta_t)p(S_t \mid \Theta_t)p(E_t \mid S_t) \\ &= \sum_p (\Theta_t \mid \Theta_{t-1})p(\Theta_{t-1} \mid E_{t-1})p(S_t \mid \Theta_t)p(E_t \mid S_t), \end{aligned} \tag{4}$$

where $p(\Theta_t \mid \Theta_{t-1})$ is the state transition probability between two consecutive time slices and $p(\Theta_{t-1} \mid E_{t-1})$ is the posterior at the preceding time slice and the prior at the current time. The probabilities in (3) and (4) can be fulfilled directly by efficient BN probabilistic inference algorithms [49]. The transitional probability function should be a time-variant function that the transitional probabilities decay gracefully as the time between two consecutive slices increases.

5.4 Active Fusion of Facial Features

A facial expression involves simultaneous changes of facial features on multiple facial regions. In addition, some of the facial deformations extracted from face images possibly result from the errors in facial feature detection and tracking due to the limitation of tracking accuracy. Emotional states vary over time in an image sequence and so do the facial visual cues. For facial activities at a given time, there is a subset of visual information that is the most informative for the current goal and that maximally reduces the ambiguity of classification. If we can actively and purposefully choose such visual cues for fusion, we can achieve a desirable result in a timely and efficient manner while reducing the ambiguity of classification to a minimum.

From the multisensory information fusion point of view, we have n hypothesis of expression states, $\Theta = \{\theta_1, \cdots, \theta_n\}$. The visual observations, $\mathbf{E} = \{E_1, \cdots, E_m\}$, which is obtained from m diverse visual sources, forms an information vector. The information fusion is to estimate a posterior probability that $\Theta = \theta_i$ is true given $\mathbf{E}$, i.e., $P(\Theta = \theta_i \mid \mathbf{E})$. By the most informative sources, we mean the sensory data from a subset, $E \subset \mathbf{E}$, that, after integrating with the existing data, can maximize the certainty of Θ, given E, i.e., it can lead the probability of hypothesis, $P(\Theta = \theta_i \mid E \subset \mathbf{E})$, closest to either 1 or 0, and the ambiguity of the hypothesis can be reduced to a minimum. In our approach, a fusion system is cast in a DBN framework, as shown in Fig. 11. The DBNs provide a dynamic knowledge representation and control structure that allows sensory information to be combined according to the rules of probability theory. The active sensor controller (see Fig. 11) serves as sensor cueing that allows the recognition system to actively select a subset of facial features to produce the visual information that is the most relevant to the current expression state.

Which subset of facial feature regions needs to be sensed is determined by evaluating the uncertainty reducing potential of possible consequences resulted by sensing various facial feature regions (visual channels). The uncertainty reducing potential is formulated in the framework of mutual information theory. Let us assume that we have information sources, $\mathbf{E} = \{E_1, \cdots, E_m\}$, which is a set of measurements taken from m feature regions. Let Θ be the hypothesis to confirm. Let θ be a possible outcome of Θ and e_i be a possible outcome of E_i. Furthermore, $\{E_1, \cdots, E_m\}$ have probability distributions $p(e_1)$, $p(e_2)$, $\cdots$, $p(e_m)$, respectively. According to Shannon's measure of entropy [50], the average residual uncertainty in Θ, summed over all possible outcomes e_i of E_i, can be obtained by

$$H(\Theta \mid E_i) = -\sum_{E_i,\Theta} p(\theta, e_i) \log p(\theta \mid e_i), \tag{5}$$

where $H(\cdot)$ and $H(\cdot \mid \cdot)$ denote the entropy and the conditional entropy, respectively. If we subtract $H(\Theta \mid E_i)$ from the original uncertainty in Θ prior to sensing E_i, we can obtain the total uncertainty reducing potential, I given E_i:

$$\begin{aligned} I(\Theta; E_i) &= H(\Theta) - H(\Theta \mid E_i) \\ &= -\sum_\Theta p(\theta) \log p(\theta) + \sum_{E_i,\Theta} p(\theta \mid e_i)p(e_i) \log p(\theta \mid e_i). \end{aligned} \tag{6}$$

We can extend (6) to $I(\Theta; E_1, E_2)$ for fusing two information sources E_1 and E_2, i.e.,

$$\begin{aligned} I(\Theta; E_1, E_2) &= H(\Theta) - H(\Theta \mid E_1, E_2) \\ &= -\sum_\Theta p(\theta) \log(p(\theta)) \\ &\quad + \sum_{\Theta, E_1, E_2} \Big\{ p(e_1, e_2)p(\theta \mid e_1, e_2) \log p(\theta \mid e_1, e_2) \Big\}, \end{aligned} \tag{7}$$

Fig. 12. Facial feature tracking in the odd field images sampled from IR illumination-based camera under slight head rotation, where the tracked feature points are highlighted by white dots.

where $p(\theta \mid e_1, e_2)$ is estimated by applying Bayes' theorem:

$$p(\theta \mid e_1, e_2) = \frac{p(\theta, e_1, e_2)}{\sum_\Theta p(\theta, e_1, e_2)} = \frac{p(\theta \mid e_1)p(\theta \mid e_2)}{p(\theta)\sum_\Theta \frac{p(\theta|e_1)p(\theta|e_2)}{p(\theta)}}, \quad (8)$$

and $p(e_1, e_2)$ in (7) can be estimated through $p(\theta \mid e_1, e_2)$ in (8) to get

$$p(e_1, e_2) = \frac{p(\theta, e_1, e_2)}{p(\theta \mid e_1, e_2)} = \frac{p(\theta \mid e_1)p(\theta \mid e_2)p(e_1)p(e_2)}{p(\theta)p(\theta \mid e_1, e_2)}. \quad (9)$$

In the above equations, $p(\theta \mid e_1)$ and $p(\theta \mid e_2)$ are directly obtained by BN probability inference. The prior probability of hypothesis $p(\theta)$ at time t is revised based on sensory observations from the preceding time slice $t-1$ through the state evolution probability. Since it is very rare that more than four facial regions (visual channels) have their facial feature change simultaneously, in our implementation, we allow two most informative visual channels to be fused at each time slice.

6 Experimental Results

This section presents experimental results demonstrating the promise of the proposed approach. The experiments focus on evaluating the effectiveness of our approach in modeling dynamic behavior of facial expressions for both modeling momentary intensity of facial expressions and for recognizing spontaneous and natural facial expressions. In addition, the experiments also evaluate the performance of our technique under various adverse conditions including facial feature missing due to either occlusion or head rotation and erroneous facial feature detection. Finally, experiments were also performed to demonstrate the benefits of active sensing strategy we proposed. For all experiments, we assume that there is no prior information favoring any six facial expressions available.

The hardware components of our system consist of an IR-camera, an IR illuminator installed in front of the camera, and a video decoder. The video decoder synchronizes the IR light with the camera so that we can detect dark and bright pupils. Software is developed to implement the eye tracking and facial feature tracking algorithms. For facial expression modeling and recognition, we use Intel's Probabilistic Networks Library (PNL) to build the DBN facial expression model and to perform inference. Finally, the facial expression model was interfaced with our facial feature tracking software to perform automatic facial expression recognition.

6.1 Performance in Facial Feature Tracking

This experiment evaluates the accuracy of our facial feature tracking technique. The accuracy of pupil-based facial feature tracking is evaluated by comparing the distances of two fiducial points with the manually generated ground truth; these distances include DB, CC', FC, $F'C'$, FP, $F'P'$, JK, $J'K'$, HF, $H'F'$, IG, $I'G'$, IF, $I'F'$ (see Fig. 3 for definitions). Fig. 12 gives an example of our facial feature tracking under slight head rotation, while Fig. 13 plots the errors of feature tracking with respect to the ground truth generated by manual detection. There are two major sources of error observed from Fig. 13. One source is due to foreshortening when the head rotates. Since 3D face pose (pan, tilt, swing) can be estimated from the tracked features, we can use the pose information to correct the feature displacements to eliminate feature displacement distortions due to the foreshortening effect. Another major source of error is caused by the feature occlusion due to significant head rotation. In our implementation, if the upper face features are occluded over a certain degree of significance (determined by 3D face pose), then these features on the occluded face side are no longer useful. Overall, most of tracking errors are within 3 pixels, which is sufficiently accurate to capture the facial deformation. The largest error (4-5 pixels) occurs on the mouth horizontal span (CC'). Nevertheless, this error would not affect the measure of change on CC' since the deformation on CC' itself is usually much larger due to facial expression change. For the recognition result, see Fig. 15 in the next section.

6.2 Performance in Modeling Facial Expressions

The following set of experiments is used primarily for demonstrating performance characteristics of this approach, including the aspects of modeling momentary intensity of

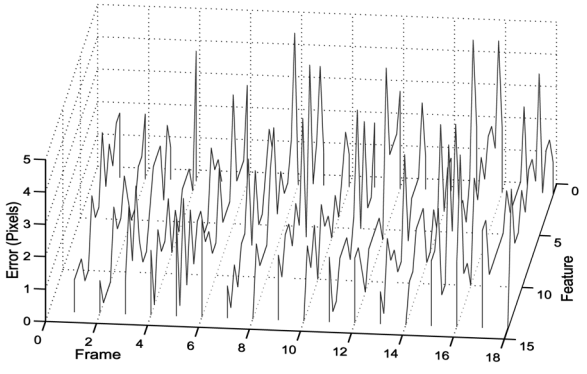


Fig. 13. The pixel errors of facial feature tracking compared with the ground truth generated by manually detection. The labels 1 to 14 on the feature axis, respectively, represent the distance of $DB, CC', FC, F'C', FP, F'P', JK, J'K', HF, H'F', IG, I'G', IF, I'F'$ (see Fig. 3 for definitions).

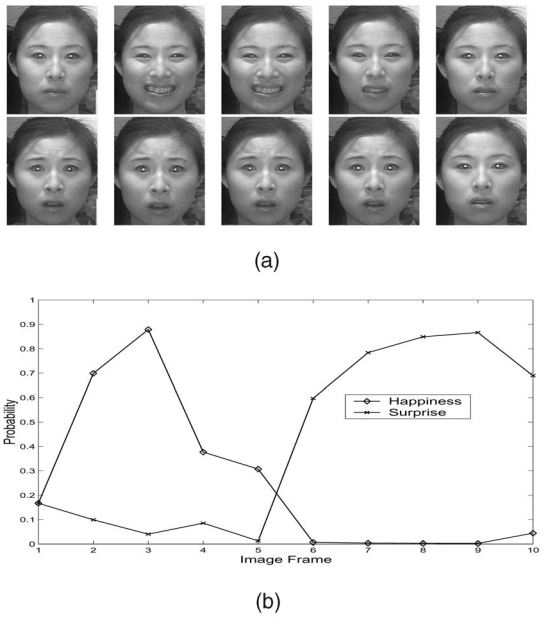


Fig. 14. (a) An image sequence shows a subject performing a smile followed by a surprise. (b) The probability distributions of facial expressions, where only the distributions of a smile and a surprise are shown.

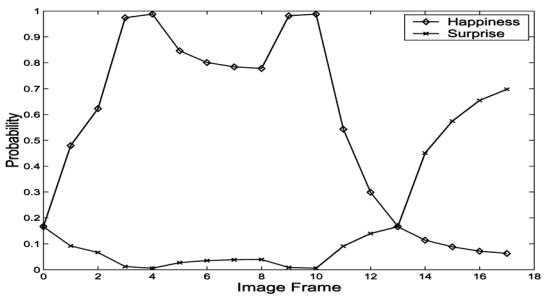


Fig. 15. The probability distributions of a smile expression and a surprise expression for the image sequence given in Fig. 12.

facial expression, handling occluded or missing facial features, and the validity of facial expression model.

We first create a short image sequence involving multiple expressions as shown in Fig. 14a. It can be seen visually that the temporal evolution of the expressions varies over time, exhibiting the spontaneous behavior. Fig. 14b provides the analysis result by our facial expression model. The result naturally profiles the momentary emotional intensity and the dynamic behavior of facial expression that the magnitude of facial expression gradually evolves over time, as shown in Fig. 14a. Such a dynamic aspect of facial expression modeling can more realistically reflect the evolution of a spontaneous expression starting from a neutral state to the apex and then gradually releasing. Fig. 15 is another example for the sequence given in Fig. 12. Since there are interpersonal variations with respect to the amplitudes of facial actions, it is often difficult to determine the absolute emotional intensity of a given subject through machine extraction. In this approach, the belief of the current hypothesis of emotional expression is inferred relying on the combined information of current visual cues through causal dependencies in the current time slice, as well as the preceding evidences through temporal dependencies. Hence, as we can observe from the results, the relative change of the emotional magnitude can

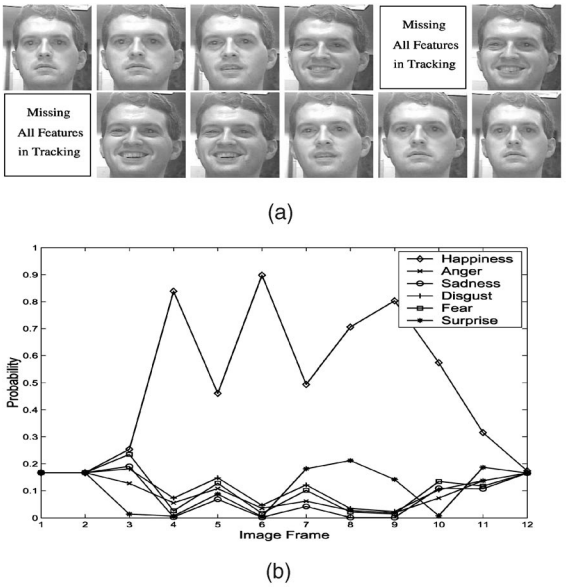


Fig. 16. (a) An image sequence assuming that the facial features in some image frames are fully missing. (b) The probability of the six facial expressions from our facial expression model. The valleys of the probability of happiness on frames 5 and 7 are caused by the absence of tacked facial features. However, the facial expression is recognized via temporal reasoning.

be well modeled at each stage of the emotional development; this is exactly what we want to achieve.

The benefits of our approach can be best shown when an image sequence presents the facial features which are misdetected or mistracked due to occlusions and image noises. Fig. 16a gives such an image sequence, in which the facial features in some frames are assumed to be completely missing. Fig. 16b depicts the result by our facial expression modeling, which provides a visual aid to see how missing sensory data are handled. From the plot in Fig. 16b, we can readily observe that, though the image sequence has facial features fully mistracked in some frames, the facial expression can still be assessed correctly by reasoning over time. The inability of current facial expression recognition systems to correlate and reason about facial temporal information over time is an impediment to provide a coherent overview of the dynamic behavior of facial expressions in an image sequence since it is often the temporal changes that provide critical information about what we try to infer and understand the human emotional expressions. However, with this approach, we can not only well-handle missing sensory data but, more importantly, reason over time as well. By integrating the current multiple visual cues and the preceding evidences, our approach tends to be more robust in handling partial occluded facial expressions. The example given in Fig. 17 further reinforces the above points. The result in Fig. 17b shows that, when major facial features are occluded, the facial expression and its intensity can still be deduced through other available features. While it will decrease the accuracy of our approach, it will not, however, totally fail our approach since our technique can still use other features and integrate this information over time in order to minimize the ambiguity caused by the occlusion. The effect of a bearded face shall be similar to that of a face occluded by a hand as shown in Fig. 17.

The correctness of our facial expression model is also evaluated by using the same image set as that used by [25], as shown in Fig. 18. Here, we take this image set as an image

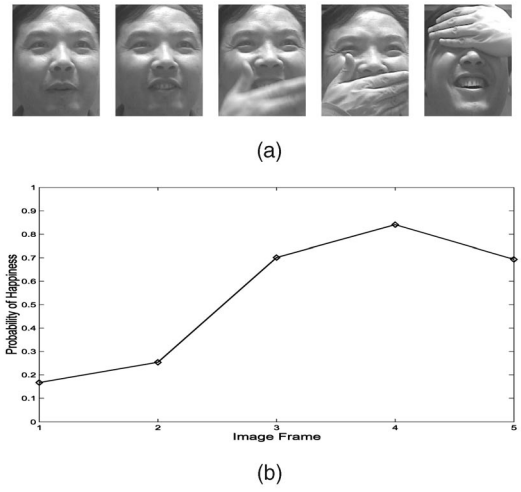


Fig. 17. (a) A posed image sequence performing occluded facial expression. (b) The result from our facial expression model.

Fig. 18. An image sequence consists of six posed facial expressions starting with a neutral expression (adapted from [25]).

sequence showing that a subject poses different expressions starting from neutral states. Notice that, for this none-IR image sequence, we manually identify the pupil positions and our facial feature detection algorithm then detects and tracks the remaining features. Table 5 gives the numerical results of probability distributions over six expressions. For a neutral facial expression (row 1 and 2 of Table 5), the probabilities over six expressions are equally likely. We can readily see from the result that, except for frame 19 which is incorrectly classified due to extreme ambiguity in its appearance, the expressions in the rest of image frames are correctly classified. The similar result was presented in [25]. In comparison, our approach emphasizes on the dynamics of facial expressions rather than the recognition accuracy of individual face images.

To evaluate the proposed active sensing strategy, we use a segment of an image sequence, as shown in Fig. 19a, for illustration. As we have previously remarked, the facial deformations extracted from images may involve the measurement errors caused by feature mistracking. If we fuse all current available visual cues at a time, the feature changes caused by the measurement errors will also be included and, consequently, cause more ambiguities in classification. Additionally, if we gather all visual channels for the information of facial feature changes each time, this will cost unnecessary computations. We therefore choose actively and purposefully the most informative visual cues for fusion. In this example, we select the two most informative visual channels to acquire the visual evidences for integration, based on the uncertainty reducing potential and the availability of facial feature changes in that channel as discussed in Section 5.4. The comparative result given in Fig. 19b shows that, in terms of uncertainty reduction, the active fusion is better than the passive fusion (fusing all available visual cues in this case). The importance of active fusion rests on its dual role as both reducing the ambiguity of classification and increasing efficiency.

TABLE 5
Classification Results for the Image Sequence in Fig. 18

Frame	Happiness	Sadness	Surprise	Anger	Disgust	Fear
1	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
2	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
3	**0.7000**	0.0021	0.0055	0.0012	0.2669	0.0242
4	**0.8921**	0.0090	0.0022	0.0070	0.0781	0.0117
5	**0.9474**	0.0006	0.0015	0.0134	0.0307	0.0064
6	0.0617	**0.7120**	0.0748	0.0003	0.0281	0.1230
7	0.0057	**0.8715**	0.0372	0.0001	0.0078	0.0777
8	0.0034	**0.9072**	0.0270	0.0001	0.0064	0.0560
9	0.0022	0.4384	**0.4793**	0.0047	0.0025	0.0728
10	0.0008	0.3046	**0.6513**	0.0004	0.0035	0.0394
11	0.0009	0.2001	**0.7650**	0.0004	0.0042	0.0295
12	0.0190	0.0711	0.0886	**0.7069**	0.0050	0.1094
13	0.0379	0.0635	0.0713	**0.7295**	0.0081	0.0897
14	0.0462	0.0602	0.0636	**0.7416**	0.0085	0.0798
15	0.1816	0.1598	0.0112	0.0075	**0.3924**	0.2474
16	0.1695	0.1204	0.0021	0.0005	**0.4733**	0.2342
17	0.0226	**0.4534**	0.0011	0.0001	0.3203	0.2026
18	0.0055	0.0822	0.0030	0.0121	0.2182	**0.6789**

6.3 Experiment in Image Sequences

Finally, we present a long image sequence acquired from our IR illumination-based camera system. The facial feature points are tracked by our pupil-based face tracking. The original image sequence has 600 frames containing the six emotional expressions plus the neutral states among them. It is difficult for us to demonstrate the experiment with all 600 image frames in print. We therefore provide selected images that will, hopefully, convey our results. Fig. 20 provides such a facial expression sequence of only 60 images resulting from sampling the original sequence in every 10 frames.

Fig. 21 visually plots the probability distributions over six facial expressions for the image sequence, as given in Fig. 20. Again, for the neutral state, the distribution of six expressions is equally likely. Fig. 21 shows that, as we initially expect, the overall performance of our facial expression understanding system based on modeling dynamic behavior of facial expressions is excellent. Nevertheless, a few of image frames are wrongly classified, e.g., the 54th and the 55th frame. In these two instances, as illustrated in Fig. 22, only the deformation on the lower face (Jaw Drop) is detected, while it fails to capture the subtle changes on the upper face. In fact, the 55th frame is also confusing even by manual inspection. Now, we take the results from the sixth and the 33rd frame in Fig. 21 to further illustrate the significance of modeling dynamic behavior of facial expressions. Based on the data extracted from the sixth frame, the face actually acts as AU10 (Upper Lip Raiser), which is a primary AU of a disgust. However,

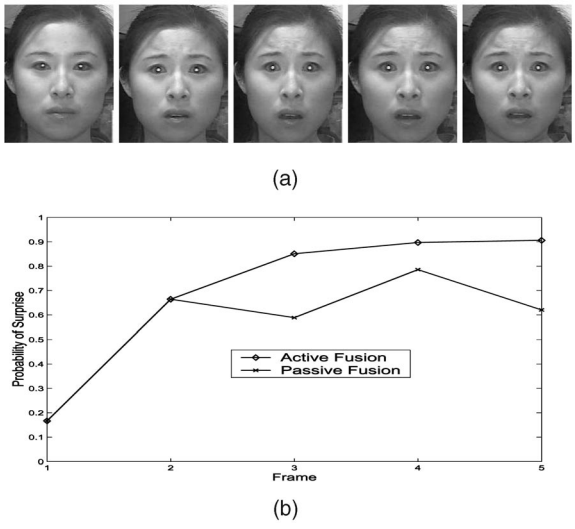


Fig. 19. (a) A segment of an image sequence. (b) Comparative result between active and purposeful fusion of visual cues and passive fusion of visual cues (fusing all available visual cues in this example).

correlating with evidences from previous time slice, the face image in the sixth frame is assessed as happiness (probability 0.3991) rather than a disgust (probability 0.3446). It can be observed from Fig. 20 that the subject of this image performs a very light smile and, thus, the classification agrees with our visual inspection. Take the 33rd frame for another instance. There are no facial deformations detected in this frame according to the feature extraction data. Nevertheless, the facial expression of this frame is correctly inferred by correlating with previous evidences. The results consistently confirm that, as indicated by the theory, the dynamic aspect of our approach can lead to be more robust for facial expression analysis in image sequences.

To further quantitatively characterize the performance of our technique, we study the facial expression recognition performance with respect to manual recognition. We conducted this experiment using the 600-frame sequence. Those manually labeled frames are then compared with the classification by our automated system. Table 6 summarizes the statistics of wrong classifications from the 600-frame sequence compared against the ground truth by the manual inspection. The table gives the number of wrongly classified image frames, where the facial expressions in the first row are the ground-truth expressions while the facial expression in the first column are the recognized facial expressions. This table quantitatively provides the classification errors of our system. The performance of our system is apparently very good.

7 Discussion and Conclusion

In this paper, we present a new approach to spontaneous facial expression understanding in image sequences. We focus our attention not only on the nature of the deformation of facial features, but also on their temporal evolution with human emotions. The problem of modeling dynamic behavior of facial expression in image sequences falls naturally within the framework of our theory on multisensory information fusion with DBNs. A particularly interesting aspect of this approach rests on its dual role as both modeling the dynamic behavior of facial expression and performing active multisensory visual information fusion. The accuracy and robustness of our approach lies in the systematic modeling of the uncertainties and dynamics of facial behavior, as well as in the integration of evidences both spatially and temporally. The efficiency of our approach is achieved through active sensing by dynamically selecting the most informative sensing sources for integration.

Accurate facial feature tracking is important for correct facial expression interpretation. Our system will perform poorly or even fail if facial features are not accurately tracked. This may happen when the user experiences a sudden head movement or a significant external illumination source saturates a portion of the face image. In addition, while our system can tolerate certain degrees of occlusion as demonstrated by our experiments, significant occlusion of some important facial features for an extended period of time can lead to the failure of our system.

In addition, there are three limitations in our approach to facial expression analysis. First, the accuracy of our facial feature detection and tracking relies on an IR-illumination camera system at the expense of the additional hardware. This may limit the application of our system to nonmobile people. Second, since feature displacements are measured with respect to their neutral positions, the knowledge about someone's facial features in the neutral facial expression has to be acquired prior to analyzing facial expressions. Third, a large amount data is necessary to accurately parameterize the framework.

Compared with the existing works on facial expression analysis, our approach enjoys several favorable properties:

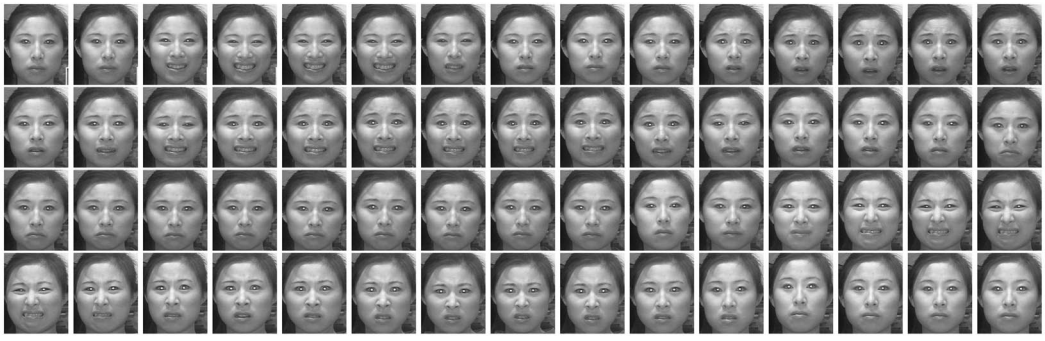


Fig. 20. An image sequence sampled from our IR illumination-based camera system (the sequence is from left to right and top to bottom).

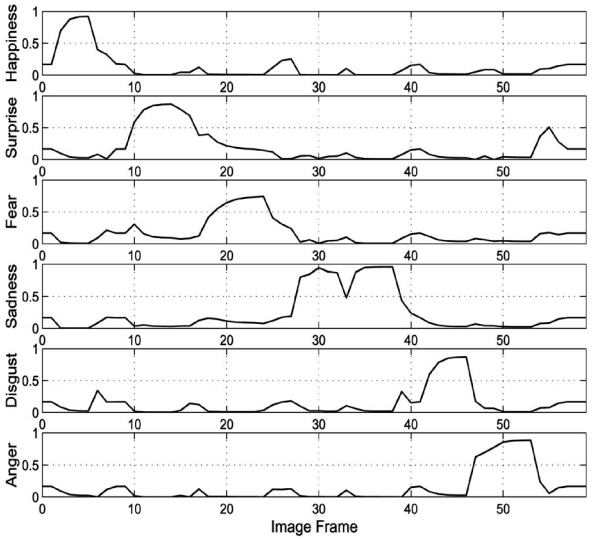


Fig. 21. The dynamic behavior and momentary intensity of facial expressions interpreted by our facial expression model. The image sequence is given in Fig. 20.

1. It has expressive power to capture the dependencies, uncertainties, and temporal behaviors exhibited by facial expressions, so that dynamic behaviors of facial expressions can be well-modeled.
2. It allows actively selecting a subset of the most relevant facial visual cues at a current time to correlate with previous visual evidences, so that the uncertainty of classification can be reduced to a minimum at each time.
3. Taking advantage of probabilistic semantics of DBNs and multisensory information fusion, our approach is more robust for facial expression understanding and for handling occluded facial expressions.
4. It allows an image sequence to have multiple expressions and two different expressions do not require to be temporally segmented by a neutral state, so that the facial expression analysis becomes more flexible in the dynamic imagery.

Finally, we would like to mention that, while our system is based on FACS, facial animation parameters (FAPs) adopted by MPEG-4 standard [51] offer an attractive alternative. Unlike FACS, which are facial muscle-based, FAPs represent facial expressions using facial feature points, which may prove to be more convenient for vision-based facial expression representation. Current FAPs, however, do not include transient facial features such as wrinkles and furrows, which are crucial in identifying facial actions.

(a) (b)

Fig. 22. An illustration of wrong classification in Fig. 20: (a) the 54th frame and (b) the 55th frame.

TABLE 6
The Statistics of Wrong Classification from the 600-Frame Sequence

	Smile	Sadness	Surprise	Anger	Disgust	Fear	Neutral
Smile							
Sadness							4
Surprise				4			
Anger			6				
disgust		2					
Fear				3			
Neutral							

Furthermore, FAPs are primarily designed for facial expression animation instead of for recognition. Nevertheless, further effort is warranted in comparing the two schemes and in studying the applicability of FAPs for vision-based facial expression representation and recognition.

ACKNOWLEDGMENTS

The authors gratefully acknowledge constructive comments from the anonymous reviewers that significantly improved the presentation of this paper. The authors would like to thank Michael J. Lyons for providing the results of Fig. 18 for their comparison. Thanks should also go to Zhiwei Zhu for providing numerous image sequences used for this research. This material is based upon work supported by a grant from the US Army Research Office under Grant No. DAAD19-01-1-0402, with which Qiang Ji is the principal investigator.

REFERENCES

[1] M.S. Bartlett, B. Braathen, G.L. Littlewort-Ford, J. Hershey, J. Fasel, T. Mark, E. Smith, T.J. Sejnowski, and J.R. Movellan, "Automatic Analysis of Spontaneous Facial Behavior: A Final Project Report," Technical Report MPLab-TR2001.08, Univ. of California at San Diego, Dec. 2001.

[2] P. Ekman and W.V. Friesen, *Facial Action Coding System (FACS): Manual.* Palo Alto, Calif: Consulting Psychologists Press, 1978.

[3] M. Kato, I. So, Y. Hishnuma, O. Nakamura, and T. Minami, "Description and Synthesis of Facial Expressions Based on Isodensity Maps," *Visual Computing,* T. Kunii, ed., pp. 39-56, 1991.

[4] G.W. Cottrell and J. Metcalfe, "Face, Emotion, Gender Recognition Using Holos," *Advances in NIPS,* R. P. Lippman, ed., pp. 564-71, 1991.

[5] A. Rahardja, A. Sowmya, and W.H. Wilson, "A Neural Network Approach to Component versus Holistic Recognition of Facial Expressions in Images," *Proc. SPIE, Intelligent Robots and Computer Vision X: Algorithms and Techniques,* vol. 1607, pp. 62-70, 1991.

[6] H. Kobayashi and F. Hara, "Recognition of Six Basic Facial Expressions and Their Strength by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 381-386, 1992.

[7] G.D. Kearney and S. McKenzie, "Machine Interpretation of Emotion: Design of Memory-Based Expert System for Interpreting Facial Expressions in Terms of Signaled Emotions (JANUS)," *Cognitive Science,* vol. 17, no. 4, pp. 589-622, 1993.

[8] H. Ushida, T. Takagi, and T. Yamaguchi, "Recognition of Facial Expressions Using Conceptual Fuzzy Sets," *Proc. IEEE Int'l Conf. Fuzzy Systems,* pp. 594-599, 1993.

[9] K. Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Trans.,* vol. E74, no. 10, pp. 3474-3483, 1991.

[10] Y. Yacoob and L. Davis, "Recognition Facial Expressions by Spatio-Temporal Analysis," *Proc. Int'l Conf. Pattern Recognition,* pp. 747-749, 1994.

[11] M. Rosenblum, Y. Yacoob, and L. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," *Proc. IEEE Workshop Motion of Non-Rigid and Articulated Objects,* pp. 43-49, 1994.

[12] M. Pantic and L. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[13] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 974-989, Oct. 1999.

[14] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition,* no. 36, pp. 259-275, 2003.

[15] Y. Yacoob and L.S. Davis, "Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 6, pp. 636-642, June 1996.

[16] I.A. Essa and A.P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 757-763, July 1997.

[17] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Difference in Face Expression," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 396-401, 1998.

[18] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequence Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 6, pp. 569-579, June 1993.

[19] N.M. Thalmann, P. Kalra, and M. Escher, "Face to Virtual Face," *Proc. IEEE,* vol. 86, no. 5, pp. 870-883, 1998.

[20] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 743-756, July 1997.

[21] M. Pantic and L. Rothkrantz, "Expert System for Automatic Analysis of Facial Expression," *J. Image and Vision Computing,* vol. 18, no. 11, pp. 881-905, 2000.

[22] Y. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 97-115, Feb. 2001.

[23] S.M. Bartlett, P.A. Viola, T.J. Sejnowski, B.A. Golomb, J. Larsen, J.C. Hager, and P. Ekman, "Classifying Facial Action," *Advances in Neural Information Processing Systems 8,* D. Touretzki, M. Mozer, and M. Hasselmo, eds., pp. 823-829, 1996.

[24] C. Padgett and G. Cottrell, "Representing Face Images for Emotion Classification," *Advances in Neural Information Processing Systems,* M. Mozer, M. Jordan, and T. Petsche, eds., vol. 9, 1997.

[25] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1357-1362, Dec. 1999.

[26] P.N. Belhumeor, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Project," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 6, pp. 711-720, June 1997.

[27] C. Huang and Y. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action," *J. Visual Comm. and Image Representation,* vol. 8, no. 3, pp. 278-290, 1997.

[28] Z. Zhu, Q. Ji, K. Fujimura, and K. Lee, "Combining Kalman Filtering and Mean Shift for Real Time Eye Tracking under Active IR Illumination," *Proc. Int'l Conf. Pattern Recognition,* Aug. 2002.

[29] J. Zhao and G. Kearney, "Classifying Facial Emotions by Backpropagation Neural Networks with Fuzzy Inputs," *Proc. Int'l Conf. Neural Information Processing,* pp. 454-457, 1996.

[30] Z. Zhang, M. Lyons, M. Schuster, and S. Akamastsu, "Comparison Between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perception," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 454-459, 1998.

[31] A. Colmenarez, B. Frey, and T.S. Huang, "A Probabilistic Framework for Embedded Face and Facial Expression Recognition," *Proc. Int'l Conf. Computer Vision and Pattern Recognition,* 1999.

[32] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Area of the Face," *J. Personality and Social Psychology,* vol. 37, pp. 2049-2059, 1979.

[33] M. Rosenblum, Y. Yacoob, and L.S. Davis, "Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture," *IEEE Trans. Neural Networks,* vol. 7, no. 5, pp. 1121-1137, 1996.

[34] M.J. Black and Y. Yacoob, "Recognizing Facial Expression in Image Sequences Using Local Parameterized Models of Image Motion," *Int'l J. Computer Vision,* vol. 25, no. 1, pp. 23-48, 1997.

[35] N. Oliver, A. Pentland, and F. Bérard, "LAFTER: Lips and Face Real Time Tracker with Facial Expression Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1997.

[36] J.J. Lien, T. Kanade, J.F. Cohn, and C. Li, "Detection, Tracking, and Classification of Action Units in Facial Expression," *J. Robotics and Autonomous Systems,* vol. 31, pp. 131-146, 1997.

[37] Y. Zhang and Q. Ji, "Facial Expression Understanding in Image Sequences Using Dynamic and Active Visual Information Fusion," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* 2003.

[38] F. Pighin, R. Szeliski, and D. Salesin, "Modeling and Animating Realistic Faces from Images," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 143-169, 2002.

[39] H. Tao and T. Huang, "Visual Estimation and Compression of Facial Motion Parameters: Elements of a 3D Model-Based Video Coding System," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 111-125, 2002.

[40] S. Goldenstein, C. Vogler, and D. Metaxas, "Statistical Cue Integration in DAG Deformable Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 7, pp. 801-813, July 2003.

[41] C. Morimoto, D. Koons, A. Amir, and M. Flicker, "Framerate Pupil Detector and Gaze Tracker," *Proc. IEEE Int'l Conf. Computer Vision Frame-Rate Workshop,* Sept. 1999.

[42] P.S. Maybeck, *Stochastic Models, Estimation, and Control.* Academic Press, Inc., 1979.

[43] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.

[44] L.G. Farkas, *Anthropometry of the Head and Face.* New York: Raven Press, 1994.

[45] L. Wiskott, J.-M. Fellous, N. Kruger, and C.V. Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 775-779, July 1997.

[46] P. Ekman, "Facial Expressions of Emotion: An Old Controversy and New Findings," *Philosophical Trans. Royal Soc. London,* vol. B, no. 335, pp. 63-69, 1992.

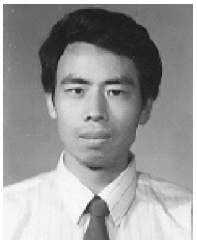
[47] V.I. Pavlovic, "Dynamic Bayesian Networks for Information Fusion with Applications to Human-Computer Interfaces," PhD thesis, Univ. of Illinois at Urbana-Champaign, 1999.

[48] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, 1989.

[49] J. Pearl, *Probability Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, Calif.: Morgan Kaufmann, 1988.

[50] C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical J.,* vol. 27, pp. 379-423, 1948.

[51] MPEG, "ISO/IEC 14496-MPEG-4 International Standard," 1998.



Yongmian Zhang received the PhD degree in computer engineering from the University of Nevada, Reno. He is currently a research fellow in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. His areas of research include information fusion, computer vision, probability modeling, and embedded and real-time systems. He is a member of the IEEE.

Qiang Ji received the PhD degree 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. His areas of research include computer vision, probabilistic reasoning for decision making and information fusion, pattern recognition, and robotics. He 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. He is a senior member of the IEEE.