A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding

Yan Tong, *Member*, *IEEE*, Jixu Chen, *Student Member*, *IEEE*, and Qiang Ji, *Senior Member*, *IEEE*

Abstract—Facial expression is a natural and powerful means of human communication. Recognizing spontaneous facial actions, however, is very challenging due to subtle facial deformation, frequent head movements, and ambiguous and uncertain facial motion measurements. Because of these challenges, current research in facial expression recognition is limited to posed expressions and often in frontal view. A spontaneous facial expression is characterized by rigid head movements and nonrigid facial muscular movements. More importantly, it is the coherent and consistent spatiotemporal interactions among rigid and nonrigid facial motions that produce a meaningful facial expression. Recognizing this fact, we introduce a unified probabilistic facial action model based on the Dynamic Bayesian network (DBN) to simultaneously and coherently represent rigid and nonrigid facial motions, their spatiotemporal dependencies, and their image measurements. Advanced machine learning methods are introduced to learn the model based on both training data and subjective prior knowledge. Given the model and the measurements of facial motions, facial action recognition is accomplished through probabilistic inference by systematically integrating visual measurements with the facial action model. Experiments show that compared to the state-of-the-art techniques, the proposed system yields significant improvements in recognizing both rigid and nonrigid facial motions, especially for spontaneous facial expressions.

Index Terms—Facial action unit recognition, face pose estimation, facial action analysis, facial action coding system, Bayesian networks.

1 Introduction

Facial action is one of the most important sources of information for understanding emotional state and intention [1]. Spontaneous facial behavior is characterized by rigid head movement, nonrigid facial muscular movements, and their interactions. Rigid head movement characterizes the overall 3D head pose, including rotation and translation. Nonrigid facial muscular movement results from the contraction of facial muscles and characterizes the local facial action at a finer level. The Facial Action Coding System (FACS) developed by Ekman and Friesen [2] is the most commonly used system for facial behavior analysis. Based on FACS, nonrigid facial muscular movement can be described by a set of facial action units (AUs), each of which is anatomically related to the contraction of a specific set of facial muscles.

An objective and noninvasive system for facial action understanding has applications in human behavior science, human-computer interaction, security, interactive games, computer-based learning, entertainment, telecommunication, and psychiatry. However, developing such a system faces several challenges:

- First, facial actions are rich and complex. Thousands of distinct nonrigid facial muscular movements (different AU combinations) have been observed so far [3] and most of them differ subtly in a few facial features.
- Second, compared to the highly controlled conditions of posed facial expressions, spontaneous facial expressions often co-occur with natural head movement when people communicate with others. For example, a person may express his agreement by nodding his head and smiling simultaneously. Hence, faces sometimes are partially occluded in some images. This makes it more challenging for accurately measuring facial motions.
- Third, most of the spontaneous facial expressions are activated without significant facial appearance changes, that is the amplitudes of the spontaneous facial expressions are smaller than those of the posed facial expressions. In addition, the spontaneous facial expression often has a slower onset phase compared to the posed facial expression [4].
- Fourth, the spontaneous facial expression may have multiple apexes and the expression does not always follow a neutral-expression-neural temporal pattern [5] as for the posed facial expressions. Moreover, multiple facial expressions often occur sequentially.
- Fifth, the subtle facial deformations and frequent head movements make it more difficult to label the facial expression data. Hence, human labeling is difficult and less reliable.

- *Y. Tong is with the Visualization and Computer Vision Lab, GE Global Research Center, One Research Circle, KW-C410 Niskayuna, NY 12308. E-mail: tongyan@ge.com.*
- *J. Chen and Q. Ji are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590. E-mail: {chenj4, jiq}@rpi.edu.*

Manuscript received 18 Nov. 2007; revised 19 June 2008; accepted 21 Nov. 2008; published online 4 Dec. 2008.
Recommended for acceptance by K. Murphy.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-11-0778.
Digital Object Identifier no. 10.1109/TPAMI.2008.293.

Due to these challenges, individually recognizing facial actions is not accurate and reliable for spontaneous facial expressions. Hence, understanding spontaneous facial action requires not only improving facial motion measurements, but, more importantly, requires exploiting the spatiotemporal interactions among facial motions since it is these coherent, coordinated, and synchronized interactions of facial motions that produce a meaningful facial display. By explicitly modeling and employing the spatiotemporal relationships in a facial action, the impact of these inaccurate or even erroneous facial motion measurements on facial action recognition can be minimized. In addition, even if some facial motion measurements are missing due to occlusion, they can be inferred through their associations with other facial motions. Thus, the performance of facial action recognition can be improved.

Previous research on facial action unit recognition [6] shows that AU recognition benefits from explicitly modeling the relationships among AUs. However, the work is limited to AU recognition on nearly frontal-view faces and ignores the impact of head movement on AU measurements. Furthermore, [6] focuses on modeling the semantic/spatial relationships among AUs and recognizes AUs from posed facial expressions. This research therefore differs from [6] in both theory and applications. Theoretically, this research models both the spatial and temporal interactions among AUs as well as modeling the interactions between the rigid motion (head pose) and the nonrigid motions (AUs). In application, this research focuses on recognizing spontaneous facial expressions, which is much more challenging than recognizing posed facial actions.

In this paper, we introduce a probabilistic facial action model based on the Dynamic Bayesian network (DBN) to simultaneously and coherently represent rigid head movement, nonrigid facial muscular movements, their spatiotemporal interactions, and their image observations in a spontaneous facial behavior. Advanced learning techniques are employed to construct the framework from both subjective knowledge and training data. Facial action recognition is accomplished through probabilistic inference by systematically integrating the facial motion measurements with the facial action model.

The proposed facial action recognition system consists of two main stages: offline facial action model construction and online facial motion measurement and inference. Specifically, using training data and subjective domain knowledge, the facial action model is constructed offline. During online recognition, as shown in Fig. 1, various computer vision techniques are used to obtain measurements of both rigid (head pose) and nonrigid facial motions (AUs). These measurements are then used as evidence by the facial action model for inferring the true states of the head pose and the AUs simultaneously. Currently, we only model left-right head movement since this type of head movement affects the AU measurement the most significantly compared to up-down and in-plane rotation and it appears frequently in spontaneous expressions. The system can be generalized to model the full range of head movement without changing the structure of the proposed facial activity model. The experiments show that, compared to the state-of-the-art techniques, the facial action recognition with the proposed system is improved significantly, especially for spontaneous facial expressions.

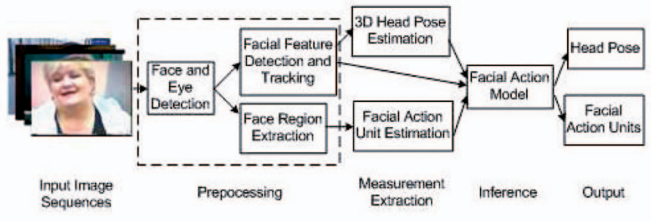


Fig. 1. The flowchart of the online facial action recognition system.

2 Related Work on Facial Action Recognition

Over the past 15 years, there has been extensive research in computer vision on recognizing facial actions. Detailed surveys of previous work can be found in [7], [8], [9], [10], [1], [11]. However, most of the previous research is limited to either estimating head poses on neutral faces or recognizing facial expressions/AUs on nearly frontal-view faces due to the challenges mentioned previously. Since, in spontaneous facial action, people usually express emotions with head movement, the assumption of nearly frontal view is not realistic.

Most recently, research has been focused on improving facial action recognition under realistic conditions, including spontaneous facial expressions with varying head pose. Recent work on facial action analysis can be classified into three groups based on how these methods deal with the relationships between the head movement and nonrigid facial muscular movements. The first group of methods [12], [13] explicitly represents and recognizes the facial expression on 3D facial expression databases. The second group [14], [15], [16], [17], [18], [19] assumes that the 3D head pose is independent of nonrigid facial muscular movements and estimates the 3D pose and nonrigid facial motions sequentially and separately. The third group [20], [21], [22], [23], [24], [32] explicitly models the coupling between rigid head movement and nonrigid facial motions for facial action analysis and simultaneously recovers the rigid and nonrigid facial motions. In contrast to recognizing facial action from deliberate facial display, automatic understanding of spontaneous facial action has been of greater interest in recent years [25], [26], [4], [27], [15], [28], [29], [30], [31], [32]. In this section, we present a brief review of previous approaches to facial action analysis.

2.1 3D Facial Expression Recognition

Since nonrigid facial muscular movements produce 3D shape deformation of the facial surface, the first group of methods [12], [13] extracts the 3D facial geometrical deformation caused by facial expression changes and recognizes the facial expression from facial deformations extracted from a 3D image database. Wang et al. [12] propose a 3D primitive surface feature based on principal curvature analysis on a 3D triangularized facial mesh. Facial expression recognition is performed on the static 3D

images, whereas the dynamic nature of facial action is ignored. Instead of employing only the static 3D images, Chang et al. [13] use 3D facial geometrical deformation for facial expression recognition from 3D videos.

Three-dimension-based facial expression recognition is not affected by illumination changes and is independent of pose variation. However, these methods require high-quality capture of texture and 3D geometrical data and thus are not applicable to many applications due to excessive hardware requirement. Furthermore, the computational complexity of the 3D methods on the dense data is considerable.

2.2 Sequential and Separate Extraction of Rigid and Nonrigid Facial Motions

Assuming that 3D head pose is independent of nonrigid facial muscular movements, the second group of methods [14], [15], [16], [17], [18], [19] estimates 3D pose and nonrigid facial motions sequentially and separately in two steps. Usually, a tracking process is performed first and the 3D pose is estimated from tracked salient facial feature points. Next, the facial expression is recognized from the pose-free facial texture or from the extracted nonrigid facial motions after eliminating the effect of pose.

Since rigid motion and nonrigid motions are nonlinearly coupled in the projected 2D facial shape/appearance, head pose estimation is not reliable under varying facial expressions. Likewise, because it is difficult to isolate the motion caused by facial expression from that caused by head movement, facial expression recognition is not accurate under varying head pose.

2.3 Simultaneous Recovery of Rigid and Nonrigid Facial Motions

The coupling between rigid head movement and nonrigid facial motions can be modeled explicitly. The methods described in [20], [21], [22], [24] estimate rigid head movement and nonrigid facial muscular movements simultaneously based on a 3D face model. The interaction between 3D head pose and facial expression is explicitly modeled as a nonlinear function.

Marks et al. [24] model the image sequence as a stochastic process generated by object motion, including both rigid and nonrigid motion, object texture, and background texture. They track object motion, object texture, and background texture simultaneously by probabilistic filtering. Although they recover the 3D head pose and the nonrigid facial deformation, they do not perform facial expression or facial action recognition.

Vasilescu and Terzopoulos [23] propose a tensor face model based on multilinear image analysis, where the face image is generated from several modes such as identity of subject, head pose, and facial expression. The multilinear image analysis assumes that each mode is allowed to vary in turn, while the remaining factors are fixed. Hence, the head pose parameters and the facial expressions are explicitly modeled by two modes in the tensor representation, respectively. The core tensor represents the interaction between the modes.

Based on an Active Appearance Model, Lucey et al. [32] separate the rigid and nonrigid motions by dividing the shape parameters into "similarity parameters" characterizing the global head movement, and "object-specific parameters" representing the nonrigid facial transformations. The AUs are recognized from a "similarity normalized" shape or appearance. However, it requires a specific AAM for each subject, which is not realistic in many applications.

Although the above methods successfully decouple the rigid and nonrigid facial motions, head movement and facial expression are recognized independently from the recovered rigid and nonrigid motions separately. These approaches neglect the interactions between rigid and nonrigid motions. Their measurements are therefore less robust, especially for spontaneous facial expressions.

2.4 Facial Action Recognition from Spontaneous Facial Display

Most of the previous approaches recognize facial action from posed or deliberate facial displays. They are of limited practical use since only the spontaneous facial display can reflect the "true" emotion [27]. Moreover, posed facial display differs from spontaneous facial display in many aspects such as the magnitudes, dynamic properties, and the interactions with head movement since they are initiated by two distinct areas of the brain [1]. Therefore, understanding spontaneous facial display is desirable and important for many real-world circumstances. Hence, facial action analysis recently began to focus on facial action recognition from spontaneous facial behaviors rather than from posed facial behaviors. These approaches include recognizing spontaneous facial emotions/expressions [27], [31], [30] and recognizing facial AUs [25], [26], [4], [15], [28], [29], [32].

Unfortunately, besides performing the facial action recognition on different facial expression data (spontaneous facial expression data versus posed facial expression data), most of the existing methods for recognizing spontaneous facial expressions employ the same techniques as for posed facial action, without exploiting the specific properties of spontaneous facial action. Therefore, these methods have the following limitations.

First, some of the current methods recognize each AU individually. However, since spontaneous facial action often produces subtle facial appearance changes rather than exaggerated appearance changes, recognizing AUs at low intensity levels is extremely difficult for current machine vision techniques. Therefore, the recognition accuracy on the spontaneous facial display degrades significantly compared to that on the posed facial display. In addition, for spontaneous facial expressions, AUs often occur in combination, and the appearance of an AU in a combination may be different from its appearance when occurring alone. This nonadditive effect makes it more difficult to recognize AUs individually.

Second, most of the current methods ignore the dynamic characteristics of AUs, which include the dynamic development of each AU and the dynamic relationships among AUs. However, recent psychological studies [33], [34], [35] show that the dynamic characteristics are very important to interpreting naturalistic human behavior. Valstar et al. [28] perform AU recognition for the eyebrow movement, and find that spontaneous eyebrow motion can be explicitly distinguished from posed eyebrow motion by employing

the dynamic properties of the related AUs such as the activating speed, magnitude, and the occurrence order of AUs. Moreover, the research by Cohn and Schmidt [4] shows that spontaneous smile usually has a relatively slower and smoother onset, and that the intensity of lip corner motion is a strong linear function of time in contrast to the posed smile.

Finally, in spontaneous facial displays, facial expression changes are often accompanied with natural head movements. Understanding spontaneous facial action should, therefore, deal with the large facial shape/appearance variations caused by both head movement and nonrigid facial muscular movements. Although most of the existing methods try to separate the facial motions caused by facial muscular movements from those caused by head movement either manually [25], [15] or automatically [26], [27], [30], [31], they generally ignore the interactions between the head movement and facial muscular movements.

In summary, current work focuses on either recognizing one type of facial motion while ignoring the other, or recognizing both motions separately while ignoring their interactions. Hence, these approaches cannot recognize facial actions reliably and robustly, especially for spontaneous facial expressions. In contrast, this work explicitly models the spatiotemporal interactions among rigid and nonrigid facial motions and uses the model to improve the facial expression recognition.

3 Facial Action Modeling

3.1 Overview of the Facial Action Model

A spontaneous facial action consists of rigid head motion, nonrigid facial deformations, and their interactions. In the scenario of facial action analysis from 2D images, the 2D facial shape ($\mathbf{S}_{2D}$) can be generated by a stochastic process involving three hidden causes we want to infer: head pose, 3D facial shape, and nonrigid facial muscular movements. The 3D facial shape ($\mathbf{S}_{3D}$) characterizes the intrinsic property of a subject, and it differs from subject to subject. The nonrigid facial muscular movements can be systematically represented by a set of AUs. The AUs are anatomically related to the contraction of the facial muscles as defined in [36] and cause the 3D shape deformation of the facial surface. Hereafter, we use $\mathbf{AU}$ to represent a set of AUs of interest (the nonrigid facial muscular movements). The head pose denoted by $Pose$ characterizes the overall head movement and causes the changes in the position and shape of the 2D face on the image. In addition, through various computer vision techniques, we can obtain measurements for these hidden causes.

Based on these causal relationships, we propose using a Bayesian network (BN) to model $\mathbf{S}_{3D}$, $\mathbf{AU}$, $Pose$, $\mathbf{S}_{2D}$, and to capture their relationships, as shown in Fig. 2a. A BN is a directed acyclic graph (DAG), where each node represents a random variable and the link between two variables characterizes the causal relationship between them. Such a model is capable of representing the statistical dependencies among $Pose$, $\mathbf{AU}$, and their interactions through $\mathbf{S}_{2D}$. Furthermore, the nodes in a BN can be grouped into hidden nodes and measurement nodes. $\mathbf{S}_{3D}$, $Pose$, $\mathbf{S}_{2D}$, and $\mathbf{AU}$ are modeled as hidden nodes. Their true states are not directly observable, and only inferred from the corresponding image measurements through the model. Hence, we associate each hidden node with a measurement node, shown as the shaded nodes in Fig. 2a. Hereafter, the shaded nodes denote the measurement nodes, and the unshaded nodes represent the hidden nodes. The measurement nodes denoted by $\mathbf{O}_{S_{3D}}$, $\mathbf{O}_{Pose}$, $\mathbf{O}_{S_{2D}}$, and $\mathbf{O}_{AU}$, respectively, represent the measurements of the corresponding hidden nodes, and their states are obtained through some computer vision techniques.

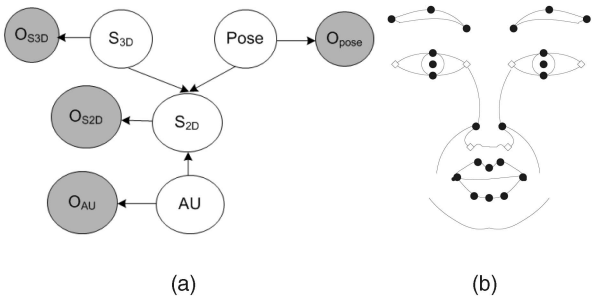


Fig. 2. (a) A graphical model to represent the causal relationships among elements of a facial action, where the shaded nodes represent the measurements of the connected hidden nodes. (b) Facial feature points on a frontal-view face: The black dots represent the local feature points, whereas the white diamonds represent the global feature points.

Given the model, facial action recognition is to find the optimal states of $Pose$ and $\mathbf{AU}$ by maximizing the joint probability of $Pose$ and $\mathbf{AU}$ given their measurements as follows:

$$Pose^*, \mathbf{AU}^* = \underset{Pose,\mathbf{AU}}{\operatorname{argmax}}\, p(Pose, \mathbf{AU} | \mathbf{O}_{Pose}, \mathbf{O}_{AU}, \mathbf{O}_{S_{3D}}, \mathbf{O}_{S_{2D}}). \quad (1)$$

In the following sections, we gradually show how the relationships in Fig. 2a can be expanded and enriched, based on which we can solve for (1).

3.2 Modeling Rigid Motion with $2D$ Global Shape

In this research, $\mathbf{S}_{3D}$ is represented by a vector of 28 facial feature points, which are located around each facial component (e.g., mouth, eye, nose, etc.), as shown in Fig. 2b. Consequently, the projection of $\mathbf{S}_{3D}$ on the 2D image plane is a 2D shape vector ($\mathbf{S}_{2D}$) containing the corresponding 28 2D points.

Given a 3D face, the deformation of $\mathbf{S}_{2D}$ reflects the action of both $Pose$ and $\mathbf{AU}$. Specifically, $Pose$ and $\mathbf{AU}$ may affect different sets of 2D facial feature points. Based on this understanding, we have defined two types of feature points: *2D global feature points* (e.g., the points on the lower nose corners and the points at eye corners) as represented by the white diamonds and *2D local feature points* (e.g., the points on the eyelids and the points on the lips) as represented by the black dots in Fig. 2b. On one hand, the 2D global feature points are relatively invariant to the AUs, and their movements are primarily caused by $Pose$. For instance, whether the eye is open or closed does not change the positions of the eye corners. On the other hand, the 2D local feature points are not only affected by $Pose$; they are also sensitive to the AUs. Consequently, $\mathbf{S}_{2D}$ can be decomposed into a global facial shape ($\mathbf{S}_{2D,g}$)

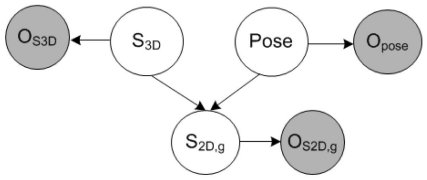


Fig. 3. $Pose$ and $\mathbf{S}_{3D}$ directly affect $\mathbf{S}_{2D,g}$. The shaded nodes are the corresponding measurement nodes.

containing the 2D global feature points and a local facial shape ($\mathbf{S}_{2D,l}$) consisting of the 2D local feature points.

$\mathbf{S}_{2D,g}$ is directly affected by $\mathbf{S}_{3D}$ and $Pose$: $\mathbf{S}_{3D}$ governs the shape of $\mathbf{S}_{2D,g}$, whereas $Pose$ controls both the position and shape of $\mathbf{S}_{2D,g}$. This causal dependency can be represented by the directed links from $Pose$ and $\mathbf{S}_{3D}$ to $\mathbf{S}_{2D,g}$, as shown in Fig. 3. Furthermore, the measurement nodes $\mathbf{O}_{S_{3D}}$, $\mathbf{O}_{Pose}$, and $\mathbf{O}_{S_{2D,g}}$ are introduced to represent the visual measurements for $\mathbf{S}_{3D}$, $Pose$, and $\mathbf{S}_{2D,g}$, respectively.

Given the causal relationships shown in Fig. 3, we need to define the states for each node and, then, learn the model parameters associated with each node. $\mathbf{S}_{3D}$ represents the unknown 3D facial shape of a subject and is characterized by a continuous 3D shape vector consisting of 28 facial feature points from neutral faces. Accordingly, its measurement $\mathbf{O}_{S_{3D}}$ is a 3D shape vector with same dimension as $\mathbf{S}_{3D}$ and represents the measured 3D facial shape. Since $\mathbf{S}_{3D}$ is hidden, its true states can only be inferred from its measurement $\mathbf{O}_{S_{3D}}$. To acquire $\mathbf{O}_{S_{3D}}$, we first perform face and eye detection on a neutral face with frontal view by using a boosted eye detection algorithm as in [37]. Once the face and eye centers are obtained, the 28 facial feature points, as shown in Fig. 2b, are detected on the neutral and frontal face similar to [38]. After that, we obtain $\mathbf{O}_{S_{3D}}$ by personalizing a trained generic 3D shape model. Specifically, the x and y-coordinates of the generic 3D shape model are adapted to current subject based on the detected positions of facial feature points on the frontal-view face. Due to the unknown depth information, the z-coordinates of the generic 3D shape model are scaled based on the size of the same frontal-view face to approximate the face depth for each individual.

Generally, three face pose angles (i.e., pan, tilt, and roll) are used to represent $Pose$. In this work, however, we only focus on modeling and recognizing the head pose variation for the pan angle (left-right rotation), which affects the shape and appearance of 2D face the most compared to the up-down and in-plane head movements. For simplicity, $Pose$ is represented by three different views: left, frontal, and right, which correspond to three discrete states ($Pose \in \{0, 1, 2\}$) in the proposed system. To obtain the measurement of $Pose$ (state of $\mathbf{O}_{Pose}$), we first estimate the three face pose angles through the weak perspective projection model based on the personalized 3D facial shape and the estimated global feature points by using a technique as described in [21]. Then, the continuous pan angle is discretized into one of the discrete states.

$\mathbf{S}_{2D,g}$ is represented by a continuous shape vector consisting of the 2D global feature points. Its measurement $\mathbf{O}_{S_{2D,g}}$ is obtained by tracking the global facial feature points in each image frame. Specifically, we use active shape models and Gabor wavelet for facial feature tracking as described in [39].

In this work, for each node without parents like $\mathbf{S}_{3D}$ and $Pose$, it is parameterized by its prior probability. For the continuous node with discrete/continuous parents like $\mathbf{S}_{2D,g}$, it is characterized by Conditional Probabilistic Distribution (CPD) defined as conditional probability of a node X, given its parents $pa(X)$, i.e., $p(X|pa(X))$, whereas, for the discrete node with discrete parents, it is characterized by Conditional Probabilistic Table (CPT) defined as $p(X|pa(X))$, similar to CPD.

The prior probabilities $p(Pose)$ and $p(\mathbf{S}_{3D})$ can be learned given the training data. The CPD of $\mathbf{S}_{2D,g}$ can be defined as we see here [40]:

$$p(\mathbf{S}_{2D,g} = s_{2D,g}|Pose = k, \mathbf{S}_{3D} = s_{3D}) = (2\pi)^{-\frac{d_g}{2}}|\Sigma_{gk}|^{-\frac{1}{2}} exp\left(-\frac{\gamma_{gk}^2}{2}\right), \quad (2)$$

where $Pose$ is at its kth state, s_{3D} is an instantiation of $\mathbf{S}_{3D}$, $s_{2D,g}$ is an instance of $\mathbf{S}_{2D,g}$, d_g is the dimension of $\mathbf{S}_{2D,g}$, and γ_{gk}^2 is defined as a Mahalanobis distance:

$$\gamma_{gk}^2 = (s_{2D,g} - \mathbf{W}_{gk} * s_{3D} - \mu_{gk})^T \Sigma_{gk}^{-1} (s_{2D,g} - \mathbf{W}_{gk} * s_{3D} - \mu_{gk}) \quad (3)$$

with the corresponding mean shape vector μ_{gk}, regression matrix $\mathbf{W}_{gk}$ and covariance matrix Σ_{gk}. Based on the conditional independence embedded in the BN, we can learn the μ_{gk}, $\mathbf{W}_{gk}$, and Σ_{gk} locally, as shown in Fig. 3, from the training data consisting of $\mathbf{S}_{2D,g}$, $\mathbf{S}_{3D}$, and $Pose$.

The link between the hidden node and measurement node represents the measurement uncertainty and is parameterized by the CPD/CPT for the measurement node. For example, $p(\mathbf{O}_{Pose}|Pose)$ represents the accuracy of the pose estimation algorithm, while $p(\mathbf{O}_{S_{2D,g}}|\mathbf{S}_{2D,g})$ models the accuracy of the facial feature tracking process.

3.3 Modeling the Relationship between $\mathbf{S}_{2D,g}$ and 2D Local Facial Component Shapes

The 2D local facial shape $\mathbf{S}_{2D,l}$ is further partitioned into four components for eyebrows, eyes, nose, and mouth, which are denoted by $\mathbf{S}_{2D,B}$, $\mathbf{S}_{2D,E}$, $\mathbf{S}_{2D,N}$, and $\mathbf{S}_{2D,M}$, respectively. The two eyes (or eyebrows) are considered as one facial component because of their symmetry. Let $\mathbf{S}_{2D,l_j}$ be any one of the 2D local facial component shape. $\mathbf{S}_{2D,l_j}$ is indirectly affected by $Pose$ through $\mathbf{S}_{2D,g}$. Given $\mathbf{S}_{2D,g}$, the position (center) of each $\mathbf{S}_{2D,l_j}$ can be estimated, independent of $Pose$. For example, the center of eye can be estimated from the eye corners, which are part of $\mathbf{S}_{2D,g}$. Hence, this causal relationship can be represented by a directed link from $\mathbf{S}_{2D,g}$ to each $\mathbf{S}_{2D,l_j}$, as illustrated in Fig. 4a.

3.4 Modeling the Relationships between AU and 2D Local Facial Component Shapes

Activating the AUs[1] produces significant changes in the shape of the facial component. For example, activating AU27 (mouth stretch) results in a widely open mouth; and

1. In this work, we intend to model and recognize a set of commonly occurring AUs including AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser and lid compressor), AU7 (lid tightener), AU9 (nose wrinkler), AU12 (lip corner puller), AU15 (lip corner depressor), AU17 (chin raiser), AU23 (lip tightener), AU24 (lip Presser), AU25 (lips part), and AU27 (mouth stretch).

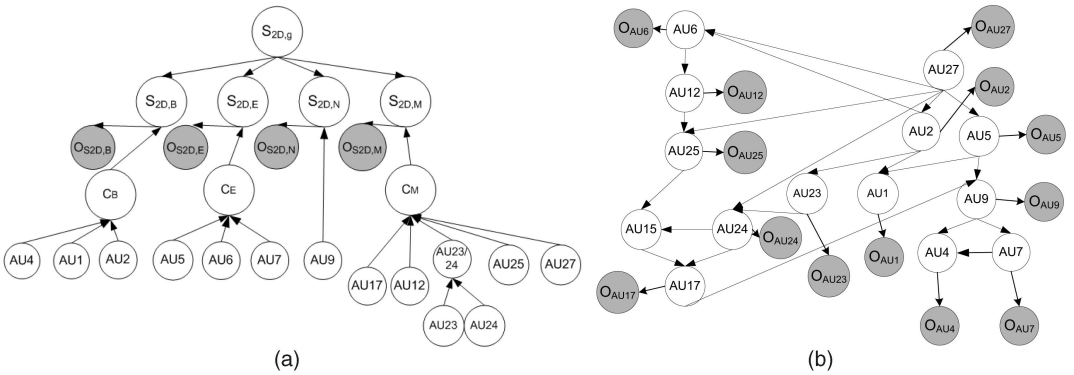


Fig. 4. (a) The relationships between 2D facial shapes and AUs. Both $\mathbf{S}_{2D,g}$ and AUs affect the 2D local facial component shapes. Intermediate nodes (C_B, C_E, and C_M) are introduced to model the correlations among AUs and to reduce the number of AU combinations to model. (b) The semantic relationships among AUs, which are crucial to produce a coherent and meaningful display. Details may be found in [6]. The shaded nodes are the corresponding measurement nodes in (a) and (b).

activating AU4 (brow lowerer) makes the eyebrows lower and pushed together. As a result, the corresponding 2D local facial component shape is also controlled by the AUs, besides the rigid head movement. For instance, there are six target AUs (AU12, AU17, AU23, AU24, AU25, and AU27) controlling mouth movements, and three target AUs (AU1, AU2, and AU4) controlling eyebrow movements. We, therefore, could directly connect the related AUs to the corresponding 2D local facial components to represent the causal relationships among them. For example, AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer) could be connected to $\mathbf{S}_{2D,B}$, while AU9 (nose wrinkler) could be connected to $\mathbf{S}_{2D,N}$ since activating AU9 will pull the infraorbital triangle upward.

However, directly connecting all related AUs to one facial component would result in too many AU combinations, most of which rarely occur in daily life. For example, based on the analysis of the training data, there are only eight common AU or AU combinations for the mouth in spite of 64 potential AU combinations. Thus, only a set of common AU or AU combinations, which produce significant nonrigid facial actions, is sufficient to control the shape variations of the facial component. As a result, a set of intermediate nodes (i.e., "C_B," "C_E," and "C_M" for eyebrow, eye, and mouth, respectively) are explicitly introduced to model the correlations among AUs and to reduce the number of AU combinations. The intermediate variables are modeled as the hidden effects from the corresponding AUs and their states are determined by the states of the corresponding AUs. Fig. 4a shows the modeling for the 2D local facial component shapes, including the relationships with the AUs, the relationships with $\mathbf{S}_{2D,g}$ as well as the relationship with their measurements.

Each AU has two discrete states which represent the "presence/absence" states of the AU. We will discuss the modeling for AUs in details in the later section. The intermediate nodes (i.e., C_B, C_E, and C_M) are discrete nodes, each state of which represents a specific AU/AU combination related to a local facial component. For instance, "C_B" has five states, each of which represents the presence of an AU or AU combination related to eyebrow movement. The CPT $p(C_i|pa(C_i))$ for each intermediate node C_i is specified based on the data analysis. For example, we assign $p(C_B = 0|AU1 = 0, AU2 = 0, AU4 = 0) = 0.9$, if the eyebrow is at the neutral state, whereas $p(C_B = 1|AU1 = 1, AU2 = 1, AU4 = 0) = 0.9$, if the eyebrow is entirely raised up.

Each $\mathbf{S}_{2D,l_j}$ (i.e., $\mathbf{S}_{2D,E}$, $\mathbf{S}_{2D,B}$, $\mathbf{S}_{2D,N}$, and $\mathbf{S}_{2D,M}$) has continuous state and is represented by a continuous shape vector containing the corresponding local feature points. Its CPD is parameterized as a Gaussian distribution. For example, for $\mathbf{S}_{2D,B}$, its CPD $p(\mathbf{S}_{2D,B} = s_{2D,B}|\mathbf{S}_{2D,g} = s_{2D,g}, C_B = k)$ is assumed to satisfy a Gaussian distribution as follows:

$$\begin{aligned} &p(\mathbf{S}_{2D,B} = s_{2D,B}|\mathbf{S}_{2D,g} = s_{2D,g}, C_B = k) \\ &\quad = (2\pi)^{-\frac{d_B}{2}}|\Sigma_{Bk}|^{-\frac{1}{2}} exp\left(-\frac{\gamma_{Bk}^2}{2}\right), \end{aligned} \tag{4}$$

where C_B is at its kth state, $s_{2D,g}$ is an instance of $\mathbf{S}_{2D,g}$, $s_{2D,B}$ is the state of $\mathbf{S}_{2D,B}$, d_B is the dimension of $\mathbf{S}_{2D,B}$, and γ_{Bk}^2 is defined as a Mahalanobis distance:

$$\begin{aligned} \gamma_{Bk}^2 = &(s_{2D,B} - \mathbf{W}_{Bk} * s_{2D,g} - \mu_{Bk})^T \Sigma_{Bk}^{-1} (s_{2D,B} \\ &- \mathbf{W}_{Bk} * s_{2D,g} - \mu_{Bk}) \end{aligned} \tag{5}$$

with corresponding mean shape vector μ_{Bk}, regression matrix $\mathbf{W}_{Bk}$, and covariance matrix Σ_{Bk}. Given the training data of $\mathbf{S}_{2D,B}$, $\mathbf{S}_{2D,g}$, and the related AUs, we can learn the parameters μ_{Bk}, $\mathbf{W}_{Bk}$, and Σ_{Bk} locally. The parameters for $\mathbf{S}_{2D,E}$, $\mathbf{S}_{2D,N}$, and $\mathbf{S}_{2D,M}$ are defined and learned likewise.

The measurement for the $2D$ local facial component shape such as $\mathbf{O}_{S_{2D,E}}$ and $\mathbf{O}_{S_{2D,B}}$ is obtained by tracking the corresponding local feature points in each image frame similar to obtain $\mathbf{O}_{S_{2D,g}}$ described above.

3.5 Semantic AU Relationships Modeling

So far, we have modeled relationships between $Pose$ and AUs through the 2D facial shapes, but have not discussed the modeling of relationships among AUs, i.e., modeling semantic dependencies among facial action units. Based on the study in [6], AUs are spatially and semantically related in order to create a meaningful facial display. In particular, there are two important semantic relationships among the AUs: co-occurrence relationships and mutually exclusive relationships. The co-occurrence relationships characterize

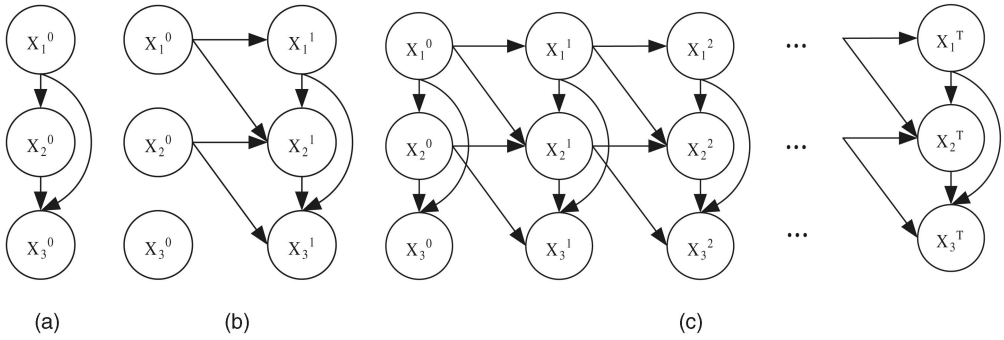


Fig. 5. A pair of (a) static network B_0 and (b) transition network $B_{\rightarrow}$ defines the dynamic dependencies for three random variables X_1, X_2, and X_3. (c) The corresponding "unrolled" DBN for $T+1$ time slices.

some groups of AUs, which usually appear together to show meaningful facial displays. For example, if the mouth and the eyes are observed to be widely open, the eyebrows are most likely raised up since it implies a surprise expression. On the other hand, based on the alternative rules provided in the FACS manual, some AUs are mutually exclusive since "it may not be possible anatomically to do both AUs simultaneously" or "the logic of FACS precludes the scoring of both AUs" [36]. For instance, the lips cannot be parted as AU25 (lips part) and pressed as AU24 (lip presser) simultaneously. These semantic relationships among AUs are especially important for understanding spontaneous facial display. For example, the enjoyment or Duchenne smiles are often accompanied by the facial muscular movements related to AU12 (lip corner puller) and AU6 (cheek raiser and lid compressor), whereas the miserable or dampened smiles are often produced by AU12 (lip corner puller) and AU15 (lip corner depressor) [41], [42]. Fig. 4b shows the model for characterizing the spatial dependencies among some AUs, where we can learn the CPTs for all AUs simultaneously. Details about this model and its parameterizations can be found in [6].

As shown in Fig. 4b, each AU is associated with a shaded measurement node denoted by $\mathbf{O}_{AU_i}$. Given the normalized face image based on the detected eye positions, we can extract the measurement for each AU through a general-purpose learning mechanism based on Gabor wavelet-based feature representation and AdaBoost classification following the work in [15]. Since we intend to recognize the AUs under varying face pose, for each AU, an AdaBoost classifier is constructed for each state of $Pose$ (frontal, left, and right), respectively. Assuming that the face pose varies smoothly over time, for each AU, the AdaBoost classifier, which corresponds to the state of $Pose$ estimated in the previous frame, is used to extract its measurement for the current frame. The CPT for each measurement node $p(O_{AU_i}|AU_i)$ is determined based on the AU recognition rate of the specific AdaBoost classifier.

4 Modeling the Dynamic Relationships among AUs

In a spontaneous facial action, AUs not only are spatially dependent on each other, but also show strong temporal dependencies to represent different naturalistic facial expressions. Nishio et al. [43] have shown that, when the mouth moves prior to the eye movement, a smile expression is mostly interpreted as a smile of enjoyment. On the contrary, when the eyes move prior to the mouth movement, it is mostly interpreted as a dampened smile.

Generally, there are two types of temporal dependencies among AUs: intradependency and interdependency. Intradependency characterizes the self-temporal evolution of an AU, while interdependency captures temporal dependencies between different AUs, i.e., an AU will be activated following the activation of another AU. For example, in a spontaneous smile, AU12 (lip corner puller) is usually first activated to express a slight emotion; then, with the increasing of emotion intensity, AU6 (cheek raiser and lid compressor) is activated in an average of 0.4 second after the activation of AU12 [42]; and after both the actions reach their apexes simultaneously, AU6 is relaxed and AU12 is gradually released before both of them return to the neutral state. Furthermore, due to the variability among individuals and different contexts, the dynamic relationships among AUs are stochastic. Systematically capturing such temporal dependencies among AUs and incorporating them into the facial action recognition process is especially important for interpretation of spontaneous facial behaviors.

4.1 A DBN for Modeling Dynamic Dependencies among AUs

We propose using a DBN to model and learn the dynamic dependencies among AUs. A DBN is a directed acyclic graphical model, which models the temporal evolution of a set of random variables $\mathbf{X}$ over time [44]. It represents a generalization of the traditional dynamic models including the Hidden Markov Models and Kalman Filtering. Let $\mathbf{X}^t$ represent a set of random variables at a discrete time slice t. A DBN is defined as $B=(G,\Theta)$, where G is the model structure, and Θ represents the model parameters, i.e., the CPDs/CPTs for all nodes. There are two assumptions in the DBN model: First, the system is first-order Markovian, i.e., $P(\mathbf{X}^{t+1}|\mathbf{X}^0,\ldots,\mathbf{X}^t)=P(\mathbf{X}^{t+1}|\mathbf{X}^t)$, and second, the process is stationary, i.e., that the transition probability $P(\mathbf{X}^{t+1}|\mathbf{X}^t)$ is the same for all t. Therefore, a DBN B can be also defined by a pair $(B_0,B_{\rightarrow})$: 1) the static network $B_0=(G_0,\Theta_0)$, as shown in Fig. 5a, captures the static distribution over all variables $\mathbf{X}^0$; and 2) the transition network $B_{\rightarrow}=(G_{\rightarrow},\Theta_{\rightarrow})$, as shown in Fig. 5b, specifies the transition probability $P(\mathbf{X}^{t+1}|\mathbf{X}^t)$ for all t in finite time slices T.

Given a DBN model, the joint probability over all variables $\mathbf{X}^0, \ldots, \mathbf{X}^T$ can be factorized by "unrolling" the DBN into an extended static BN, as shown in Fig. 5c, whose joint probability is computed as follows:

$$P(\mathbf{x}^0, \ldots, \mathbf{x}^T) = P_{B_0}(\mathbf{x}^0) \prod_{t=0}^{T-1} P_{B_\rightarrow}(\mathbf{x}^{t+1}|\mathbf{x}^t), \quad (6)$$

where $\mathbf{x}^t$ represents the sets of values taken by the random variables $\mathbf{X}$ at time t, $P_{B_0}(\mathbf{x}^0)$ captures the joint probability of all variables in the static BN B_0, and $P_{B_\rightarrow}(\mathbf{x}^{t+1}|\mathbf{x}^t)$ represents the transition probability and can be decomposed as follows based on the conditional independencies encoded in the DBN:

$$P_{B_\rightarrow}(\mathbf{x}^{t+1}|\mathbf{x}^t) = \prod_{i=1}^{N} P_{B_\rightarrow}\big(x_i^{t+1}\big|pa\big(X_i^{t+1}\big)\big), \quad (7)$$

where $pa(X_i^{t+1})$ represents the parent configuration of variable X_i^{t+1} in the transition network $B_\rightarrow$, and N represents the number of random variables in $\mathbf{X}^t$. Hereafter, $pa^j(X)$ represents the jth parent configuration of variable X in a given network structure.

We intend to use DBN to learn and model two types of temporal relationships for AUs at two adjacent time slices. The intradependency is modeled as an arc linking an AU_i node at time $t-1$ (AU_i^{t-1}) to that at time t (AU_i^t) and depicts how a single AU develops over time. The interdependency is modeled as an arc from AU_i node at time $t-1$ to AU_j at time t and represents the pairwise dynamic dependency between two different AUs.

We notice that the temporal relationships are critically depending on the temporal resolution of the image sequences, i.e., the frame rate of the recorded videos. The learned temporal relationships may not be valid on the image sequences collected under a different frame rate. Therefore, we consider the temporal relationships between two fixed-size time durations instead of the temporal relationships between two successive frames. In addition, for each AU, its temporal evolution consists of a complete temporal segment lasting from $1/4$ of a second (e.g., a blink) to several minutes (e.g., a jaw clench) as described in [5]. If we choose a single frame as one time duration, we may capture many irrelevant events, whereas if we choose many frames as a duration, the dynamic relationships may not be captured. Hence, based on an analysis of the databases we use as well as on the temporal characteristics of the AUs we intend to recognize, the time duration is empirically set as $1/6$ second in this work. This way, if AU_i appears within the tth time duration, it is counted as "presence" at t, i.e., $AU_i^t = 1$. Hereafter, we use "slice" instead of "duration" for a generalized representation in the later presentation.

4.2 Constructing the Initial DBN

In this work, each AU is represented by a binary value $[0, 1]$ representing its absence/presence status. An AU is in the "presence" status, when it is activated and is at one of the three temporal states (onset, apex, and offset), whereas it is in the "absence" status, when it is at the neutral state. Since the "presence/absence" of an AU at time t depends not only on its state in previous time slice, but also on the states of other AUs, $P(AU_j^t|AU_j^{t-1}, AU_i^{t-1})$ is used to capture the dynamic relationships between AU_i and AU_j as well as the dynamic evolution of AU_j itself. For example, the positive dependency between two AUs in adjacent slices is computed as follows:

$$P\big(AU_j^t = 1\big|AU_j^{t-1} = 1, AU_i^{t-1} = 1\big) = \frac{N_{AU_j^t + AU_j^{t-1} + AU_i^{t-1}}}{N_{AU_j^{t-1} + AU_i^{t-1}}}, \quad (8)$$

where $N_{AU_j^{t-1} + AU_i^{t-1}}$ is the total number of the events that the AU combination $(AU_j + AU_i)$ is present in the $(t-1)$th slice, regardless of the presence of other AUs, and $N_{AU_j^t + AU_j^{t-1} + AU_i^{t-1}}$ is the total number of the events that AU_j is present in the tth slice and the AU combination $(AU_j + AU_i)$ is present in the $(t-1)$th slice. The other probabilities are computed similarly.

For initialization, the intradependency and interdependency are partially learned from two posed facial expression databases: namely, the Cohn-Kanade facial expression database [45] and ISL multiview facial expression database,[2] collected by our own research group. Furthermore, in order to recognize AUs in spontaneous facial expression, we will later refine the initial dynamic relationships among AUs using databases containing natural facial expressions such as [46] and [47]. Based on the data analysis from the two databases, we can find that the statistical information, extracted from training data, is consistent with the dynamic relationships among AUs found in the psychological studies [42]. For example,

$$\begin{aligned} &P(AU_6^t = 1|AU_6^{t-1} = 0, AU_{12}^{t-1} = 1) = 0.1 \\ &> P(AU_6^t = 1|AU_6^{t-1} = 0, AU_{12}^{t-1} = 0) = 0.02, \end{aligned}$$

which means AU6 (cheek raiser and lid compressor) occurs mostly after AU12 (lip corner puller) is activated.

If the probability $P(AU_j^t = 1|AU_j^{t-1} = 0, AU_i^{t-1} = 1)$ is higher than a predefined threshold T_{up} or the probability $P(AU_j^t = 1|AU_j^{t-1} = 1, AU_i^{t-1} = 0)$ is lower than a predefined threshold T_{bottom}, we assume that there is a strong dynamic dependency between AU_i and AU_j, which can be modeled with an interslice link from AU_i^{t-1} to AU_j^t in the DBN. For example, the link from AU_{12}^{t-1} to AU_6^t represents the dynamic dependency between AU6 and AU12. This way, using the statistics extracted from the two databases, an initial transition network is manually constructed as in Fig. 6a.

4.3 Learning DBN Model

Given a set of observed data $D = \{D_1, \ldots, D_M\}$, where M is the total number of training images, we can refine the initial DBN model with a structure learning algorithm, i.e., finding a DBN structure G that best fits the observed data. For learning a DBN model, the training data D should be divided into S sequences with length M_s so that $\sum_s M_s = M$. As mentioned above, a DBN consists of two parts (B_0 and $B_\rightarrow$); therefore, we should learn both of them from the training data. To evaluate the fitness of the network, we need to define a scoring function. The score of

2. More details about ISL multiview facial expression databases and other related databases can be found at http://www.ecse.rpi.edu/homepages/cvrl/database/database.html.

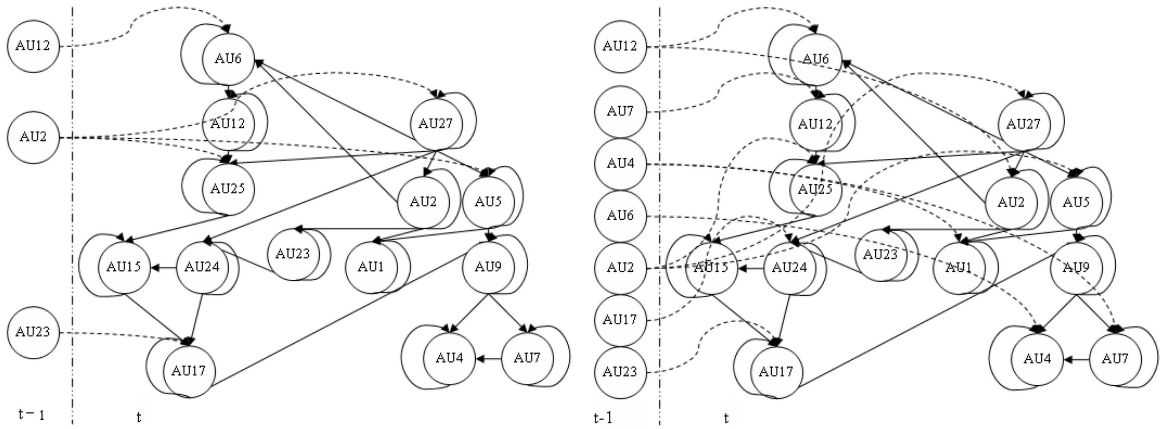


Fig. 6. (a) The initial transition network and (b) the learned transition network by the proposed algorithm for AU modeling. The self-arrow at each AU node indicates the temporal relationship of a single AU from the previous time slice to the current time slice. The dashed line with arrow from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the pairwise dynamic dependency between different AUs.

a DBN model can be defined based on the Bayesian Information Criterion (BIC) [48], i.e.,

$$Score(B) = logP(B, D) = logP(B) + logP(D|B), \quad (9)$$

where $logP(B)$ is the log prior probability of the network structure and $logP(D|B)$ is the log likelihood, which can be computed approximately as follows:

$$logP(D|B) \approx logP(D|G, \widehat{\Theta_G}) - \frac{logM}{2} Dim_G, \quad (10)$$

where the first term evaluates how well the network B fits the data D; the second term is a penalty relating to the complexity of the network; $\widehat{\Theta_G}$ is the set of parameters of G which maximizes the likelihood of the data; M is the number of training data; and Dim_G is the number of parameters.

Instead of giving an equal prior probability $P(B)$ to all possible structures, we assign a higher probability to the manually constructed initial structure B_{init} shown in Fig. 6a as in [49]. Given the scoring function, the structure learning is performed by searching the network with the highest score among the possible network structures.

Specifically, the score for a DBN model can be decomposed into two parts as follows:

$$Score(B) = Score_{B_0} + Score_{B_\rightarrow}, \quad (11)$$

where $Score_{B_0}$ and $Score_{B_\rightarrow}$ represent the score of the static network and the score of the transition network, respectively. Consequently, we can learn the structure of B_0 and the structure of $B_\rightarrow$ separately.

4.3.1 Learning the Static Network

The static network B_0 models the semantic spatial relationships among AUs within a time slice as discussed in Section 3.5. Following (9) and (10), the score for the static network is defined as follows:

$$Score_{B_0} = logP(B_0) + \sum_i \sum_j \sum_k N^0_{i,j,k} log\hat{\theta}^0_{i,j,k} - \frac{logM}{2} Dim_{B_0}, \quad (12)$$

where $\theta^0_{i,j,k}$ represents the parameters for the static network B_0, i.e., $\theta^0_{i,j,k} = P(X^0_i = k|pa^j(X^0_i))$ with i ranges over all variables, j ranges over all parent configuration of one specific variable X^0_i, and k ranges over all states of X^0_i in B_0. $N^0_{i,j,k}$ is the total number of events that X^0_i is at its kth state with jth parent configuration.

Given the definition of $Score_{B_0}$, we employ an iterated hill climbing algorithm [50] to search the optimal network structure. First, an initial network structure B_{0init} is manually constructed by combining the data analysis from the two databases (Cohn-Kanade database [45] and ISL database) and the domain knowledge from the FACS rules [36]. Then, starting from $B^0_0 = B_{0init}$, $Score_{B_0}$ is computed as in (12) for each nearest neighbor of B^0_0, which is generated from B^0_0 by adding, deleting, or reversing a single arc that is subject to the acyclicity constraint and the limitation on the upper bound of parent nodes. In this way, the network structure that has the maximum score among all of the nearest neighbors is selected as the static network B_0 as shown in Fig. 4b. Details on learning the static BN structure for modeling the semantic AU relationships may be found in [6].

4.3.2 Learning the Transition Network

Learning the transition network is more complicated than learning the static network. The transition network $B_\rightarrow$ consists of two types of links: interslice links and intraslice links. The interslice links are the dynamic links connecting the temporal variables of two successive time slices. In contrast, the intraslice links connect the variables within a single time slice, which are same as the static network structure. The score of the transition network is defined as follows:

$$Score_{B_\rightarrow} = log(P(B_\rightarrow)) + \sum_i \sum_j \sum_k N^{\rightarrow}_{i,j,k} log\hat{\theta}^{\rightarrow}_{i,j,k} - \frac{log(M-S)}{2} Dim_{B_\rightarrow}, \quad (13)$$

where $M - S$ is the total number of pairwise transitions between two successive slices in the training data and $\theta^{\rightarrow}_{i,j,k}$ represents the parameters for the transition network $B_\rightarrow$ and is defined as $\theta^{\rightarrow}_{i,j,k} = P(X^t_i = k|pa^j(X^t_i))$ for the node X^t_i

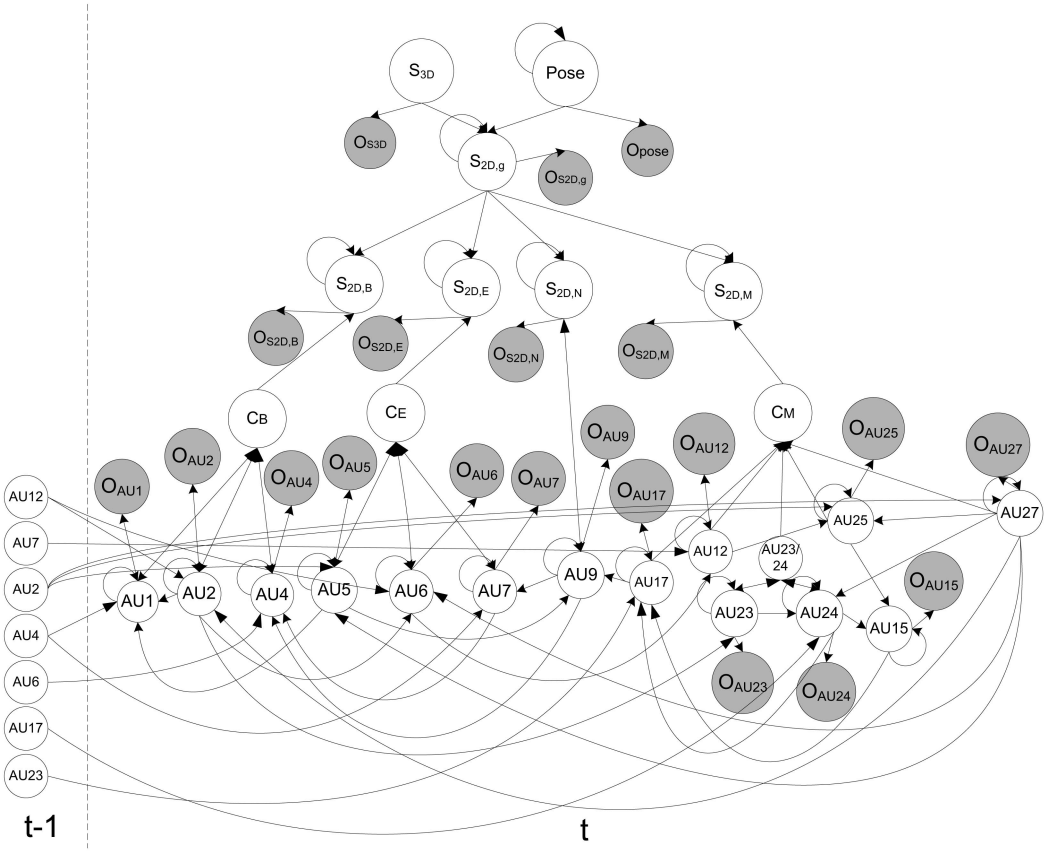


Fig. 7. The complete DBN model for facial action understanding. The shaded node indicates the observation for the connected hidden node. The self-arrow at the hidden node represents its temporal evolution from previous time slice to the current time slice. The link from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the dynamic dependency between different AUs.

at kth state given its jth parent configuration in $B_{\rightarrow}$. $N_{i,j,k}^{\rightarrow}$ accounts for the number of the instances of transitions, where X_i^t is at its kth state with jth parent configuration.

Given the definition of a score for the transition network as in (13), we need to identify a transition network structure with the highest score by a searching algorithm subject to some coherent constraints on the transition network. First, the variables $\mathbf{X}^0$, as shown in Fig. 5b do not have parents. Second, the interslice links can only have one direction, that is, from the previous time slice to the current time slice. Finally, based on the stationary assumption, both the interslice links and intraslice links should be repeated for the time slice $t \in [1, T]$. Furthermore, since the multiwise relationships are hard to capture and vary significantly from people to people, for better generalization, we only capture the strong pairwise dynamic dependencies among AUs that are true for most people and ignore the weak temporal relationships that are mostly person dependent. Therefore, an additional constraint is imposed so that each node X_i^{t+1} has at most two parents from the previous time slice t. We then apply the same hill climbing technique [50] to identify the transition network structure.

Fig. 6b shows the learned transition network by the proposed learning algorithm. Compared with the manually constructed initial transition network in Fig. 6a, the learned structure better reflects the dynamic relationships among AUs in the training data. For example, the dynamic link from AU_4^{t-1} to AU_7^t means that the eyelids intend to be narrowed by activating AU7 (lid tightener) with the increasing intensity of AU4 (brow lowerer). And the dynamic link from AU_{17}^{t-1} to AU_{24}^t means that before the lips are pressed together (AU24), it is most likely the "chin boss" [36] is already moved upward by activating AU17 (chin raiser).

5 A Complete Facial Action Model

5.1 A Comprehensive Model for Facial Action Understanding

Now we are ready to present the complete DBN model for facial action modeling, as shown in Fig. 7. Specifically, we employ the relationship between $Pose$ and $\mathbf{S}_{2D,g}$ as the global constraint for the overall system so that it will guarantee globally meaningful facial action. Meanwhile, the local structural details of the facial components not only are constrained by the local shape parameters, but are also characterized by the related AUs. In addition, the interactions between $Pose$ and AUs are indirectly modeled through $\mathbf{S}_{2D,g}$ and $2D$ local facial component shapes. Finally, the facial action measurements are systematically incorporated into the model through the shaded nodes. This model therefore completely characterizes the spatial and temporal dependencies between rigid and nonrigid facial motions and accounts for the uncertainties in facial motion measurements. Finally, we learn the transition probability for the temporal links of the DBN besides learning the static links as described in the previous sections. Specifically,

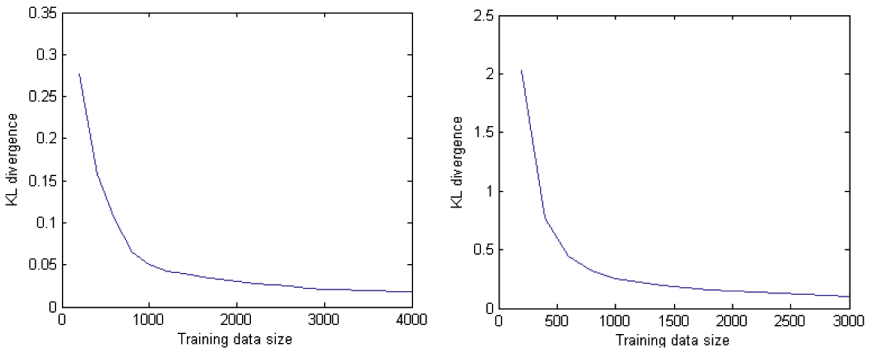


Fig. 8. The KL divergence of the parameters versus the training data size for the (a) dynamic AU model and (b) "Eye Brow" model, respectively.

based on (7), learning the transition probability associated with the temporal link is to learn the conditional probability of the temporal variable X_i, given its parent configuration in the transition network $p_{B_\rightarrow}(X_i^{t+1}|pa(X_i^{t+1}))$. Since the training data is complete, we can learn the conditional probability $p_{B_\rightarrow}(X_i^{t+1}|pa(X_i^{t+1}))$ using Maximum Likelihood (ML) estimation method.

5.2 Analysis of Training Data Size for Model Learning

As described previously, based on the conditional independencies encoded in the DBN model, the facial action model, as shown in Fig. 7, can be decomposed into a set of smaller local structures. As a result, the parameters of the facial action model can be learned individually in each local structure. Hence, fewer training data are required for learning parameters in a smaller model than are required for a large network structure. To evaluate the quantity of training data needed for learning the facial action model, we perform a sensitivity study of model learning on different amount of training data. For this purpose, the Kullback-Leibler (KL) divergences of the parameters are computed versus the number of training samples. Specifically, two local models are learned: 1) the dynamic AU model capturing the spatiotemporal relationships among all target AUs as shown in Fig. 6b; and 2) a local model called "Eye Brow" model capturing the relationships among $\mathbf{S}_{2D,g}$, $\mathbf{S}_{2D,B}$, "C_B," and the related AUs, as shown in Fig. 4a.

The experimental results reported in Fig. 8 show that, for the dynamic AU model, the learning process requires a total of only 2,000 training samples for all target AUs since all AUs have binary states. For the "Eye Brow" model, since the states of $\mathbf{S}_{2D,g}$ and $\mathbf{S}_{2D,B}$ are continuous, the training process requires 1,500 training samples. Since we have more than 2,000 training samples available in the databases, the training data for the model learning are sufficient. Furthermore, the recent studies in [51], [52] show that we can significantly reduce the training data by incorporating some qualitative knowledge in the model learning. By exploiting some qualitative constraints on the parameters, we can achieve the similar learning results with only one tenth of the training data required in the conventional learning techniques.

5.3 Facial Action Inference

Once the measurement nodes are observed, we can infer the facial action by finding the most probable explanation (MPE) of the evidence, as shown in (1). The advantage of using MPE is that it allows us to infer all the variables of interest simultaneously, instead of inferring each variable individually; therefore, it simultaneously finds the most probable state combination of $Pose$ and the AUs.

Let $Pose^t$ and $AU_{1\cdots N}^t$ represent the nodes for $Pose$ and N target AUs at time t. Given the available evidence until time t: $\mathbf{O}_{S_{3D}}, \mathbf{O}_{Pose}^{1:t}, \mathbf{O}_{S_{2D,g}}^{1:t}, \mathbf{O}_{S_{2D,l_{1\cdots L}}}^{1:t}, \mathbf{O}_{AU_{1\cdots N}}^{1:t}$, where N is the number of target AUs and L is the number of local facial component shapes, the probability $p(Pose^t, AU_{1\cdots N}^t|\mathbf{O}_{S_{3D}}, \mathbf{O}_{Pose}^{1:t}, \mathbf{O}_{S_{2D,g}}^{1:t}, \mathbf{O}_{S_{2D,l_{1\cdots L}}}^{1:t}, \mathbf{O}_{AU_{1\cdots N}}^{1:t})$ can be factorized and computed via the facial action model by performing the DBN updating process as described in [53]. Then, the true joint states of $Pose$ and the AUs are inferred simultaneously over time by maximizing $p(Pose^t, AU_{1\cdots N}^t|\mathbf{O}_{S_{3D}}, \mathbf{O}_{Pose}^{1:t}, \mathbf{O}_{S_{2D,g}}^{1:t}, \mathbf{O}_{S_{2D,l_{1\cdots L}}}^{1:t}, \mathbf{O}_{AU_{1\cdots N}}^{1:t})$. Specifically, we employ the junction tree inference algorithm in the Bayes Net Toolbox by Murphy [54] to infer the true states of $Pose^t$ and $AU_{1\cdots N}^t$.

6 Experimental Results

6.1 Facial Expression Databases

The proposed facial action analysis system is trained and tested on FACS labeled images from three databases. The first database is Cohn-Kanade DFAT-504 database [45], which consists of more than 100 subjects covering different races, ages, and genders. This database has been widely used for evaluating facial AU recognition system. However, the image sequences in Cohn-Kanade database only contain frontal-view face images. In order to demonstrate our system under more realistic circumstance, we also constructed our own database (ISL multiview facial expression database), which consists of 40 image sequences from eight subjects containing the target AUs. The ISL database is collected under uncontrolled indoor illumination and background. The subjects are instructed to perform the target AUs or the basic facial expressions (e.g., smiling or surprising) while turning around their head. Hence, the face undergoes large face pose variations (-30 degrees to 30 degrees from left to right) and significant facial expression changes simultaneously.

However, both the Cohn-Kanade database and the ISL database contain posed facial expressions, whereas the interactions between the rigid motion and nonrigid facial muscular movements and the dynamic and semantic

relationships among AUs in the spontaneous facial actions are different from those in the posed facial actions. Hence, the proposed system will be also trained and tested on spontaneous facial expression databases, whereas the training and testing procedures are the same as those on the posed facial expression databases. Therefore, we extend our work for recognizing spontaneous facial actions.

Specifically, the proposed system is trained and tested on spontaneous facial expression databases, which consists of image videos collected through three sources: 1) 37 image sequences from six videos of the Multiple Aspects of Discourse (MAD) research lab at the University of Memphis [47], 2) 10 image sequences from three videos of the Belfast natural facial expression database [46],[3] and 3) 27 image sequences of three videos obtained from the website http://www.youtube.com. In the MAD database, the subjects are providing technical assistance to some customer via headphone, while, in the Belfast database and the videos from the website, the subjects are usually being interviewed by a journalist. Hence, there are many speech-related facial motions involved in this combined spontaneous facial expression database. Furthermore, in the spontaneous database, the subjects are displaying various spontaneous facial expressions with natural head movements. The collected videos are further presegmented to only include image sequences with more active facial activity.

For this study, all of the image sequences in the databases (Cohn-Kanade database, ISL database, and spontaneous facial expression databases) are coded into AUs frame by frame. For each AU, the positive samples are chosen as the images containing the target AU at different intensity levels, and the negative samples are selected as those images without the target AU regardless the presence of other AUs. For training the facial shape models, we also manually marked the 28 feature points on some images from these databases. In the following, we will perform experimental validation on the proposed system and compare the performance of the proposed system with the state-of-the-art techniques [15], [6] on these image databases, respectively.

6.2 Evaluation on Cohn-Kanade Database

We first evaluate our system on the Cohn-Kanade database [45] for AU recognition to demonstrate the system performance on the standard database. The database is divided into eight sections, each of which contains images from different subjects. Each time, we use seven sections for training and the remaining section for testing so that the training and testing set are mutually exclusive. The average recognition performance is computed on all sections.

Fig. 9 shows the performance for generalization to novel subjects in Cohn-Kanade database of using the AdaBoost classifiers alone [15], using the semantic AU model that only models AU relationships as in [6],[4] and using the proposed model, respectively. The AdaBoost classifiers [15] achieve an average correct-positive recognition rate of 80.6 percent and an average false-positive rate of 7.84 percent for the 14 target AUs. By employing the relationships among AUs, the semantic AU method increases the average correct-positive rate to 85.8 percent and reduces the false-positive rate to

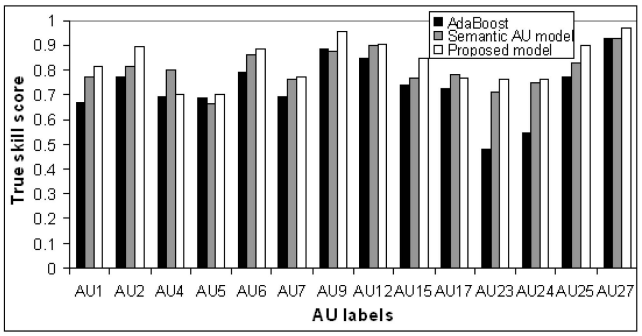


Fig. 9. Comparison of AU recognition results on the novel subjects in Cohn-Kanade database by using the AdaBoost classifier [15] (black bar), using the semantic AU model as in [6] (gray bar), and using the proposed model (white bar), respectively, based on the true skill score (Hansen Kuiper Discriminant), which is the difference between the correct-positive recognition rate and the false-positive rate.

5.54 percent. With the use of the proposed model, our system achieves an average correct-positive rate of 88.3 percent and reduces false-positive rate to 5.17 percent. Compared with the semantic AU model, the improvement by using the facial action model is not that significant. That is because the Cohn-Kanade database contains only the images with frontal-view faces and posed facial expressions. Hence, the improvement is mainly due to modeling the learned dynamic relationships among AUs. We will further demonstrate that an additional improvement can be achieved for the spontaneous facial expressions by modeling the relationships among *Pose*, AUs, and the 2D facial shapes in the latter sections.

6.3 Evaluation under Realistic Environment

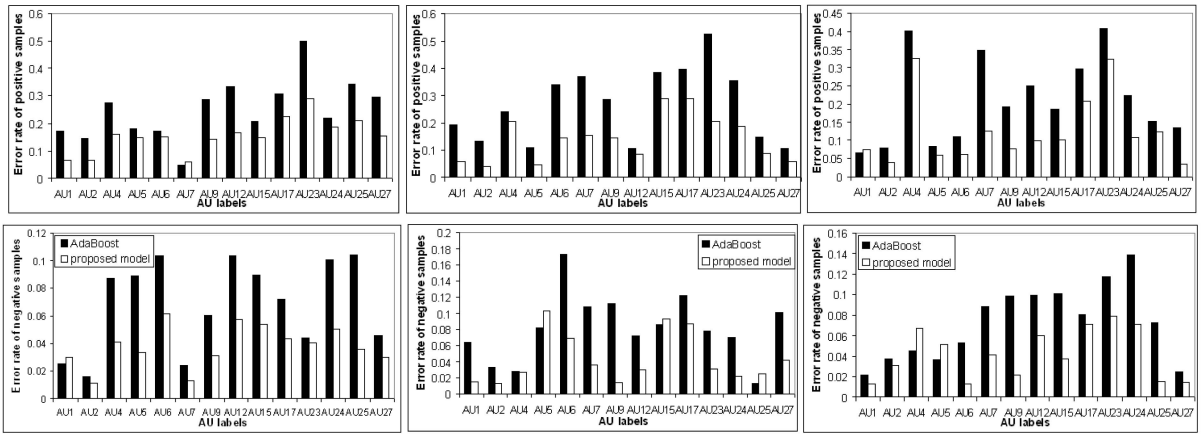
In order to demonstrate the robustness of the proposed system, we perform experiments on our own database under realistic environment where the face undergoes facial expression and face pose changes simultaneously. Fig. 10 shows an example of image sequence from the ISL database, where the subject is laughing and turning around his head.

The system is evaluated based on the leave-one-subject-out cross validation. The system performance is reported in Fig. 11. Compared to the AU recognition by the AdaBoost classifiers [15] only, we can find that using the proposed facial action model: 1) For the frontal-view face, the average relative error rate of positive samples (positive error rate or false-negative rate) decreases by 37.5 percent and the average relative error rate of negative samples (negative error rate or false-positive rate) decreases by 44.7 percent; 2) for the right-view face, the average relative positive error rate decreases by 40 percent and the average relative negative error rate decreases by 42.2 percent; and 3) for the left-view face, the average relative positive error rate decreases by 46.1 percent and the average relative negative error rate decreases by 46.8 percent. Here, the relative error rate is defined as the ratio of the error rate of the proposed method to the error rate of the AdaBoost method [15]. Especially for the AUs that are difficult to recognize, the system performance is greatly improved. For example, for AU23 (lip tighten), its positive error rate decreases from 52.3 percent to 20.5 percent, and its negative error rate decreases from 7.7 percent to 3.1 percent for the left-view face; the positive error rate of AU7 (lid tighten) decreases from 34.8 percent to 12.5 percent for the right-view face; and the positive error rate of AU6 (cheek raiser and lid compressor) decreases from 34 percent to

3. Although we obtain a total of 16 pilot videos from Belfast database, only three videos, which have minimum body movements, are employed in this work.

4. The semantic AU model focuses on modeling the semantic (static) relationships among AUs, while only two dynamic AU relationships are manually specified.

Fig. 10. An example image sequence from the ISL database where the subject is laughing with left-to-right head rotation.



(a) (b) (c)

Fig. 11. AU recognition results under realistic circumstance for (a) frontal-view faces, (b) left-view faces, and (c) right-view faces. The first row demonstrates average error rate of positive samples. The second row displays average error rate of negative samples. In each figure, the black bar denotes the result by the AdaBoost classifier [15], and the white bar represents the result by using the proposed model.

14.5 percent with a significant drop of negative error rate from 17.3 percent to 6.95 percent for the left-view face.

We also perform pose estimation on the image sequences through the probabilistic inference. As shown in Table 1, pose estimation by the proposed method is also improved compared to the pose measurement obtained by a method as in [21]. The improvement comes from modeling the interactions of *Pose* with AUs through the 2D facial shapes. In summary, the proposed system significantly improves AU recognition and pose estimation simultaneously.

6.4 Evaluation on Spontaneous Facial Expression Database

Instead of recognizing posed facial actions, it is more important to recognize spontaneous facial actions. Therefore, in the third set of the experiments, the system is trained and tested on the spontaneous facial expression databases to demonstrate the system robustness for recognizing the spontaneous facial action. Currently, the combined spontaneous facial expression database contains 74 image sequences from 13 subjects, where 63 image sequences are used for training and 11 image sequences for testing. The subjects in the training and testing sets are not exclusive since the target AUs are not present for all subjects, especially for some AUs that do not occur frequently. Fig. 12 shows an example of image sequence from the Belfast database [46] where the subject is talking under natural head movement.

TABLE 1
Comparison of Pose Estimation by Using [21] and Using the DBN Inference through the Proposed Model

view	[21]	proposed method
frontal	93%	94.3%
right	94.4%	96.1%
left	86.7%	93.4%

Since there are major differences between the spontaneous facial action and posed facial action as described in Section 1, we should learn a new facial action model using the spontaneous training data to capture the relationships among AUs, the coupling between the rigid motion and nonrigid motion, and the dynamics for spontaneous facial action. In this work, we intend to recognize 12 target AUs, which frequently occur in the spontaneous facial expression database. Fig. 13 shows the learned DBN model from the spontaneous facial expression database.

Fig. 14 shows the average AU recognition performance on the spontaneous facial expression database of using the AdaBoost classifier alone [15], using the semantic AU model as in [6], and using the proposed facial action model, respectively. The AdaBoost classifiers [15] achieve an average positive error rate of 44 percent and an average negative error rate of 8.58 percent for the 12 target AUs. By employing the relationships among AUs, the semantic AU model decreases the average positive error rate to 36.5 percent and decreases the negative error rate to 6.6 percent. With the use of the

Fig. 12. An example image sequence adapted from Belfast database [46] where the subject is talking under natural head movement.

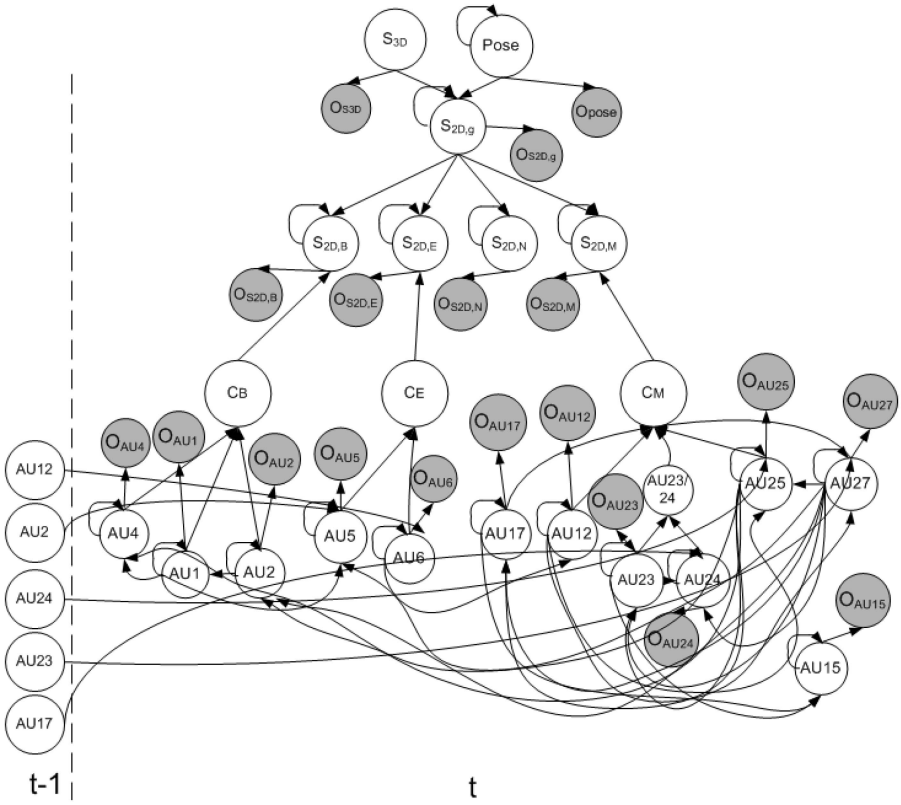
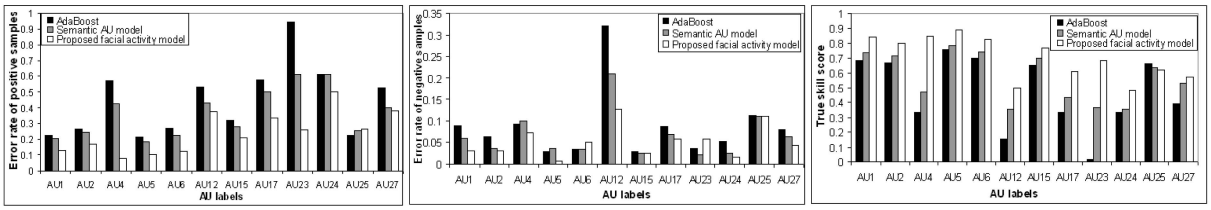


Fig. 13. The complete DBN model for spontaneous facial action recognition.



(a) (b) (c)

Fig. 14. AU recognition results on spontaneous facial expression database: (a) average positive error rates, (b) average negative error rates, and (c) average true skill scores (the difference between the correct-positive recognition rate and the false-positive rate). In each figure, the black bar denotes the result by using the AdaBoost classifier [15], the gray bar represents the result by using the semantic AU model [6], and the white bar represents the result by using the proposed facial action model.

proposed complete facial action model, the average positive error rate further decreases to 24.3 percent and the negative error rate decreases to 5.3 percent.

Compared to the recognition results from the AdaBoost classifiers [15], the system performance is greatly improved by using the proposed facial action model for some AUs. For example, the positive error rate of AU23 (lip tighten) decreases from 94.4 percent to 25.9 percent with a moderate increasing of negative error rate (from 3.6 percent to 5.8 percent), the positive error rate of AU12 (lip corner puller) decreases from 53 percent to 37.8 percent with a significant drop of negative error rate (from 32 percent to 12.8 percent), and the positive error rate of AU2 (outer brow raiser) decreases from 26.9 percent to 16.9 percent with a decreasing of negative error rate from 6.2 percent to 3.1 percent.

Furthermore, since the spontaneous facial action is often accompanied by natural head movements, only employing the relationships within AUs is not sufficient to deal with

the facial appearance variations due to varying *Pose*. By incorporating the relationships between *Pose* and AUs through the 2D facial shapes, the complete facial action model improves the system performance significantly compared to the semantic AU model [6].

7 Conclusion and Future Work

The recent research shows that the spatiotemporal relationships among the rigid and nonrigid facial motions are important for spontaneous facial action analysis and understanding. Therefore, we propose a unified facial action model based on the DBN to systematically discover and learn such relationships, and then combine them with the image observations to perform a robust and reliable recognition of spontaneous facial action. The experiments show that compared to the state-of-the-art techniques [15], [6], the proposed system yields significant improvements in both pose estimation and AU recognition, especially for the spontaneous facial expressions. The performance improvements come mainly from combining the facial action model with the facial measurements. Specifically, the erroneous AU measurements can be compensated by the model's built-in spatial and temporal relationships among AUs and the built-in relationships between rigid head motions and the nonrigid facial motions. Hence, instead of solely improving the computer vision techniques, it is important to capture the prior knowledge or context in a probabilistic manner and systematically combines the captured knowledge with the improved visual measurements to achieve a robust and accurate visual interpretation.

A binary representation for the AU state (presence/absence) is employed in this work mainly due to the limited resource of ground truth labels for AUs with multiple states (e.g., the five intensity levels defined in FACS [36]). It still remains challenging for both human and computer vision techniques to recognize multiple AU intensity levels reliably and accurately, especially for spontaneous expression. However, given sufficient and reliable AU labels with more states, the current framework can be extended without modifying the model structure.

The current system can process about 2.1 frames/second on a 2.8 GHz Pentium IV PC. It could be sped up by optimization of the code and by implementing the DBN inference in C++ instead of in Matlab. The future work will focus on the following aspects. First, we plan to extend the work to include continuous measurement of *Pose* and to include both tilt and pan rotations. Second, we will extend the DBN model to systematically model human labeling errors by introducing another layer of nodes to represent the labeling confidence. In addition, a more extensive quantitative sensitivity study on the model structure will be performed to pinpoint the exact reasons that lead to the improved performance. For applications, we would like to apply the framework to distinguish faked facial expression from genuine and natural facial expressions. The proposed method may be also applied to perform "soft" biometrics for human identification based on their facial behaviors.

Acknowledgments

Portions of the research in this paper use Cohn and Kanade's $DFAT-504$ database, Belfast natural facial expression database, and videos from Multiple Aspects of Discourse research lab at the University of Memphis. The authors gracefully acknowledge their support. This project is supported in part by the US Air Force Office of Scientific Research (AFOSR) Grant F49620-03-1-0160 and the US Defense Advanced Research Projects Agency (DARPA)/US Office of Naval Research (ONR) Grant N00014-03-1-1003.

References

[1] M. Pantic and M. Bartlett, "Machine Analysis of Facial Expressions," *Face Recognition,* K. Delac and M. Grgic, eds., pp. 377-416, I-Tech Education and Publishing, 2007.

[2] P. Ekman and W.V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.

[3] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research.* Cambridge Univ. Press, 1982.

[4] J.F. Cohn and K. Schmidt, "The Timing of Facial Motion in Posed and Spontaneous Smiles," *Int'l J. Wavelets, Multiresolution, and Information Processing,* vol. 2, pp. 1-12, Mar. 2004.

[5] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences," *IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics,* vol. 36, no. 2, pp. 433-449, Apr. 2006.

[6] Y. Tong, W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 10, pp. 1683-1699, Oct. 2007.

[7] M. Pantic and L.J.M. 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.

[8] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition,* vol. 36, no. 1, pp. 259-275, 2003.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32-80, Jan. 2001.

[10] Y. Tian, T. Kanade, and J. Cohn, "Facial Expression Analysis," *Handbook of Face Recognition,* S. Li and A. Jain, eds., Springer, 2004.

[11] M. Pantic, A. Pentland, A. Nijholt, and T.S. Huang, "Human Computing Machine Understanding of Human Behavior: A Survey," *Artificial Intelligence for Human Computing,* T.S. Huang, A. Nijholt, M. Pantic, and A. Pentland, eds., Springer Verlag, 2007.

[12] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 1399-1406, 2006.

[13] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3D Facial Expression Analysis in Videos," *Proc. Analysis and Modelling of Faces and Gestures,* pp. 293-307, 2005.

[14] B. Braathen, M.S. Bartlett, G.C. Littlewort, E. Smith, and J.R. Movellan, "An Approach to Automatic Recognition of Spontaneous Facial Actions," *Proc. Fifth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 345-350, 2002.

[15] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," *J. Multimedia,* vol. 1, no. 6, pp. 22-35, Sept. 2006.

[16] F. Dornaika and F. Davoine, "Simultaneous Facial Action Tracking and Expression Recognition Using a Particle Filter," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 1733-1738, 2005.

[17] Y. Zhang and Q. Ji, "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 699-714, May 2005.

[18] M.F. Valstar, I. Patras, and M. Pantic, "Facial Action Unit Detection Using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, Workshop Vision for Human-Computer Interaction,* June 2005.

[19] R. El Kaliouby and P. Robinson, "Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures," *Real-Time Vision for HCI,* pp. 181-200, Springer Verlag, 2005.

[20] B. Bascle and A. Blake, "Separability of Pose and Expression in Facial Tracking and Animation," *Proc. Int'l Conf. Computer Vision,* pp. 323-328, 1998.
[21] Z. Zhu and Q. Ji, "Robust Real-Time Face Pose and Facial Expression Recovery," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 681-688, 2006.
[22] M. Anisetti, V. Bellandi, E. Damiani, and F. Beverina, "3D Expressive Face Model-Based Tracking Algorithm," *Proc. Signal Processing, Pattern Recognition, and Applications,* pp. 111-116, 2006.
[23] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles: Tensorfaces," *Proc. European Conf. Computer Vision,* pp. 447-460, 2002.
[24] T.K. Marks, J. Hershey, J.C. Roddey, and J.R. Movellan, "Joint Tracking of Pose, Expression, and Texture Using Conditionally Gaussian Filters," *Advances in Neural Information Processing Systems,* vol. 17, pp. 889-896, The MIT Press, 2005.
[25] A. Kapoor, Y. Qi, and R.W. Picard, "Fully Automatic Upper Facial Action Recognition," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 195-202, 2003.
[26] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* vol. 1, pp. 610-616, 2004.
[27] N. Sebe, M. Lew, I. Cohen, S. Yafei, T. Gevers, and T. Huang, "Authentic Facial Expression Analysis," *Proc. Sixth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 517-522, 2004.
[28] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn, "Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions," *Proc. Eighth Int'l Conf. Multimodal Interfaces,* pp. 162-170, 2006.
[29] G. Littlewort, M.S. Bartlett, and K. Lee, "Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain," *Proc. Ninth Int'l Conf. Multimodal Interfaces,* pp. 15-21, 2007.
[30] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T.S. Huang, "Spontaneous Emotional Facial Expression Detection," *J. Multimedia,* vol. 1, no. 5, pp. 1-8, 2006.
[31] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion Recognition through Facial Expression Analysis Based on a Neurofuzzy Method," *Neural Networks,* vol. 18, no. 4, pp. 423-435, 2005.
[32] S. Lucey, A.B. Ashraf, and J. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," *Face Recognition Book,* K. Kurihara, ed., Pro Literatur Verlag, Apr. 2007.
[33] J. Russell and J. Fernandez-Dols, *The Psychology of Facial Expression.* Cambridge Univ. Press, 1997.
[34] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," *J. Personality and Social Psychology,* vol. 37, no. 11, pp. 2049-2058, 1979.
[35] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Oxford Univ. Press, 2005.
[36] P. Ekman, W.V. Friesen, and J.C. Hager, *Facial Action Coding System: The Manual.* Research Nexus Division, Network Information Research Corp., 2002.
[37] P. Wang and Q. Ji, "Multi-View Face and Eye Detection Using Discriminant Features," *Computer Vision and Image Understanding,* vol. 105, no. 2, pp. 99-111, Feb. 2007.
[38] Z. Zhu and Q. Ji, "Robust Pose Invariant Facial Feature Detection and Tracking in Real-Time," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 1092-1095, 2006.
[39] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji, "Robust Facial Feature Tracking under Varying Face Pose and Facial Expression," *Pattern Recognition,* vol. 40, no. 11, pp. 3195-3208, Nov. 2007.
[40] K. Murphy, "Inference and Learning in Hybrid Bayesian Networks," Technical Report CSD-98-990, Dept. of Computer Science, Univ. of California Berkeley, 1998.
[41] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage.* Norton, 1985.
[42] K. Schmidt and J. Cohn, "Dynamics of Facial Expression: Normative Characteristics and Individual Differences," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 728-731, 2001.
[43] S. Nishio, K. Koyama, and T. Nakamura, "Temporal Differences in Eye and Mouth Movements Classifying Facial Expressions of Smiles" *Proc. Third IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 206-211, Apr. 1998.
[44] T. Dean and K. Kanazawa, "Probabilistic Temporal Reasoning," *Proc. Seventh Nat'l Conf. Artificial Intelligence,* pp. 524-528, 1988.
[45] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. Fourth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 46-53, 2000.
[46] E. Douglas-Cowie, R. Cowie, and M. Schroeder, "The Description of Naturally Occurring Emotional Speech," *Proc. 15th Int'l Congress of Phonetic Sciences,* 2003.
[47] Multiple Aspects of Discourse Research Lab, http://madre searchlab.org/, 2009.
[48] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 461-464, 1978.
[49] D. Heckerman, D. Geiger, and D.M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning,* vol. 20, no. 3, pp. 197-243, 1995.
[50] S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach.* Prentice-Hall, 1995.
[51] Y. Tong and Q. Ji, "Learning Bayesian Networks with Qualitative Constraints," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2008.
[52] C.P. de Campos, Y. Tong, and Q. Ji, "Exploiting Qualitative Constraints for Learning Bayesian Network Parameters," *Proc. European Conf. Computer Vision,* 2008.
[53] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence.* Chapman and Hall/CRC, 2004.
[54] K. Murphy, "The Bayes Net Toolbox for Matlab," *Computing Science and Statistics,* vol. 33, pp. 331-350, 2001.

Yan Tong received the PhD degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, New York, in 2007. She is currently a researcher at the General Electric Global Research Center, Niskayuna, New York. Her areas of research include computer vision, pattern recognition, and human-computer interaction. She is a member of the IEEE and the IEEE Computer Soceity.

Jixu Chen received the BS and MS degrees in electrical engineering from the University of Science and Technology of China in 2003 and 2006, respectively. He is currently pursuing the PhD degree at Rensselaer Polytechnic Institute, Troy, New York. His areas of research include computer vision, pattern recognition, and their applications in human-computer interaction. He is a student member of the IEEE and the IEEE Computer Society.

Qiang Ji received the PhD degree in electrical engineering from the University of Washington. He is currently a professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He is also a program director at the US National Science Foundation (NSF), managing part of the NSF's computer vision and machine learning programs. He has also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, the Department of Computer Science at the University of Nevada at Reno, and the US Air Force Research Laboratory. He currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI. His research interests are in computer vision, pattern recognition, and probabilistic graphical models. He has published more than 150 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including the US National Science Foundation (NSF), NIH, DARPA, ONR, ARO, and AFOSR, as well as by major companies including Honda and Boeing. He is an editor for several computer vision and pattern recognition related journals and he has served as a program committee member, area chair, and program chair for numerous international conferences/workshops. He is a senior member of the IEEE and a member of the IEEE Computer Society.