Modeling and Exploiting the Spatio-temporal Facial Action Dependencies for Robust Spontaneous Facial Expression Recognition

Yan Tong
GE Global Research Center
Niskayuna, NY 12309
tongyan@research.ge.com

Jixu Chen and Qiang Ji
Rensselaer Polytechnic Institute
Troy, NY12180
qji@ecse.rpi.edu

Abstract

Facial action provides various types of messages for human communications. Recognizing spontaneous facial actions, however, is very challenging due to subtle facial deformation, frequent head movements, and ambiguous and uncertain facial motion measurements. As a result, current research in facial action recognition is limited to posed facial actions and often in frontal view.

Spontaneous facial action is characterized by rigid head movements and nonrigid facial muscular movements. More importantly, it is the spatiotemporal interactions among the rigid and nonrigid facial motions that produce a meaningful and natural facial display. Recognizing this fact, we introduce a probabilistic facial action model based on a dynamic Bayesian network (DBN) to simultaneously and coherently capture rigid and nonrigid facial motions, their spatiotemporal dependencies, and their image measurements. Advanced machine learning methods are introduced to learn the probabilistic facial action model based on both training data and prior knowledge. Facial action recognition is accomplished through probabilistic inference by systemically integrating measurements of facial motions with the facial action model. Experiments show that the proposed system yields significant improvements in recognizing spontaneous facial actions.

1. Introduction

Facial action is one of the most important sources of information for understanding emotional state and intention [14]. Spontaneous facial action consists of rigid motion, nonrigid motion, and their spatiotemporal interactions. Rigid motion characterizes the overall 3D head pose. Nonrigid motion characterizes the local facial muscular movement and can be described by 44 facial action units (AUs) based on the Facial Action Coding System (FACS) [8]. Automatically recognizing spontaneous facial action has applications in human behavior analysis, human-computer interaction, psychiatry, etc. However, due to the low intensity, nonadditive effect, and individual difference of spontaneous facial action as well as the image uncertainty, it is not accurate and reliable to recognize facial action through measuring some local aspects of facial motion individually. Hence, understanding spontaneous facial action requires not only improving facial motion measurements, but more importantly, exploiting the spatiotemporal interactions among facial motions since it is these interactions that produce a "synchronized, smooth, symmetrical, and consistent" [14] facial display. By explicitly modeling and using these inherent relationships, we can improve facial action recognition performance by compensating erroneous or missing facial motion measurements.

2. Related Work

Over the past fifteen years, there has been extensive research in computer vision on recognizing facial actions. Detailed surveys of previous work can be found in [15, 20, 14]. However, most of the previous approaches recognize facial action from posed facial displays. They are of limited practical use since only spontaneous facial display can reflect the "true" emotion [19]. Technically, current techniques have the following limitations.

First, they often recognize each AU individually. However, since spontaneous facial action often produces subtle facial appearance changes, measuring AUs at low intensity levels is not accurate and reliable. In addition, for spontaneous facial actions, AUs often occur in combination, where an AU may look different from its appearance when occurring alone. This nonadditive effect makes it more difficult to recognize AUs individually.

Second, most of them ignore the dynamic properties of AUs including the self evolution of each AU and the dynamic relationships among AUs. However, recent psychological study [2] shows that the dynamic characteristics are crucial to interpreting naturalistic human behavior. Valstar et al. [24] find that spontaneous eyebrow motion can be

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

distinguished from posed one by employing the dynamic properties of the related AUs such as the activating speed, magnitude, and the occurrence orders of AUs.

Finally, spontaneous facial expressions are often accompanied with natural head movements. Understanding spontaneous facial action should, therefore, deal with large facial shape/appearance variations caused by both rigid and nonrigid facial motions. Although the current methods try to separate rigid and nonrigid motions either manually [10, 3] or automatically [5, 19, 25, 9], they generally ignore the interactions between rigid and nonrigid motions, which are crucial for interpreting spontaneous facial behavior.

In summary, few existing methods consider the spatiotemporal interactions among facial motions. Based on our previous work [23, 21], this work explicitly models and learns the semantic and dynamic interactions among rigid and nonrigid facial motions and uses the model to improve spontaneous facial action 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 spatiotemporal interactions. In the view of facial action analysis from 2D images, 2D facial shape encodes the information of both rigid and nonrigid facial motions since it is generated from three hidden causes: head pose, 3D facial shape, and nonrigid facial muscular movements. 3D facial shape characterizes an intrinsic property of a subject. Nonrigid facial muscular movements represented by a set of AUs cause 3D shape deformation of the facial surface. Head pose produces the changes in the position and shape of the 2D face on the projected 2D image. In addition, through various computer vision techniques, we can obtain measurements for these hidden causes.

Based on the causal relationships shown in Figure 1a, we propose to use a Bayesian network (BN) to model the statistical dependencies among rigid motion, nonrigid motions, and their interactions through the 2D facial shape. Furthermore, the nodes in a BN can be grouped into hidden nodes and measurement nodes. 3D facial shape denoted by S_{3D}, facial muscular movements represented by a set of AUs, 3D head pose denoted by $Pose$, and 2D facial shape denoted by S_{2D} are modeled as hidden nodes, and their true states can be inferred from their measurements through the model. And so, we associate each hidden node with a measurement node representing the observation of the corresponding hidden node, as shown in Figure 1a.

Given the model, facial action recognition is to find the optimal states of head pose and AUs by maximizing the joint probability of pose and AUs given the measurements as follows:

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

where **AU** is the set of all target AUs; O_{pose}, $O_{\mathbf{AU}}$, $O_{S_{3D}}$, and $O_{S_{2D}}$ denote the measurements of head pose, AUs, 3D facial shape, and 2D facial shape, respectively. In the next several sections, we gradually show how the relationships in Figure 1a can be expanded and enriched.

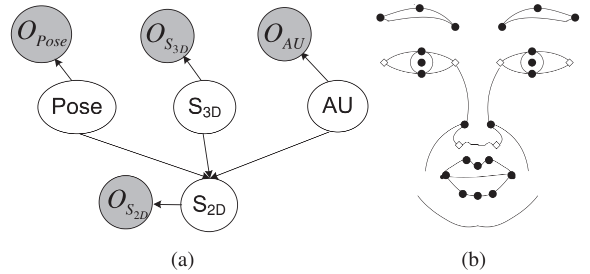


Figure 1. (a) A graphical model to represent the causal relationships among elements of a facial action, where the shaded nodes represent the measurements of the connected hidden nodes. (b) Facial feature points on a frontal view face: the black dots represent the local feature points, whereas the white dots represent the global feature points.

3.2. Modeling the Interactions Between Rigid and Nonrigid Facial Motions

In this research, the shape of a 3D face can be represented by a vector of 28 facial feature points, which are located around each facial component (e.g., mouth, eye, nose, and eyebrow), as shown in Figure 1b. Given a 3D face, the deformation of a 2D facial shape reflects the action of both head pose and facial muscular movements. Specifically, head pose and facial muscular movements may affect different sets of facial feature points. As a result, the facial feature points are further divided into global feature points, which are relatively invariant to facial muscular movements, and local feature points, which are affected by both head pose and facial muscular movements.

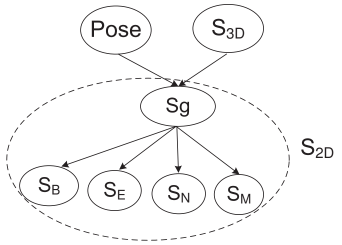


Figure 2. The head pose and 3D facial shape directly affect the 2D global shape $\mathbf{S}_g$; whereas the 2D global shape $\mathbf{S}_g$ controls the variation of each 2D local facial component shape.

(1) Modeling Rigid Motion with 2D Shape

The 2D global shape denoted by $\mathbf{S}_g$ is the projection of the 3D global feature points on the image plane. Therefore, the 3D facial shape governs the shape of the 2D global shape, whereas the 3D head pose controls both the position and shape of $\mathbf{S}_g$. This causal dependency can be represented by a directed link from head pose/3D facial shape to $\mathbf{S}_g$, as shown in Figure 2.

Furthermore, the 3D local facial shape can be partitioned into four components: eyebrows, eyes, nose, and mouth. The corresponding 2D local facial component shape is indirectly affected by the head movement through the 2D global shape $\mathbf{S}_g$. Given $\mathbf{S}_g$, the center of each 2D local facial component can be roughly estimated, independent of the head pose. For example, the center of eye can be determined, given the eye corners, which are parts of the global shape. Hence, this causal relationship can be represented by a directed link from $\mathbf{S}_g$ to each 2D local facial component shape, as shown in Figure 2.

(2) Modeling the Relationships between Nonrigid Motion and 2D Local Facial Component Shapes

The nonrigid facial muscular movements produce significant changes in the 3D shape of the facial component. These 3D facial muscular movements can be systematically represented by AUs as defined in [8]. For example, activating AU27 (mouth stretch) will produce a widely open mouth; and activating AU4 (brow lowerer) makes the eyebrows lower and pushed together.

Since the 3D shape of each facial component is determined by the related AUs, the 2D local facial component shape is controlled by the AUs, besides rigid head movement. We can model such causal relationship by directly connecting the related AUs to the corresponding facial component. For instance, AU1 (Inner brow raiser), AU2 (Outer brow raiser), and AU4 (Brow lowerer) control eyebrow movements, and can be connected to the eyebrow. However, directly connecting all related AUs to one facial component would result in too many AU combinations, most of which rarely occur in the daily life. Thus, only a set of common AUs or AU combinations 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. For example, "C_M" has 8 states, each of which represents a common AU or AU combination controlling mouth movement. Figure 3 shows the modeling of the relationships between nonrigid facial motions (AUs) and the local facial component shapes.

3.3. Modeling Semantic and Dynamic Relationships among AUs

So far, we have modeled relationships between rigid head motion and nonrigid facial motions through the 2D facial shapes, but have not discussed the modeling of spatial and temporal dependencies among AUs, which are crucial to create a meaningful and natural facial display. Tong et al [22] demonstrated that there are two important spatial semantic relationships among the AUs: co-occurrence and mutually exclusive relationships. Furthermore, AUs also show strong temporal dependencies to represent different naturalistic facial behaviors. Nishio et al. [13] 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 miserable smile.

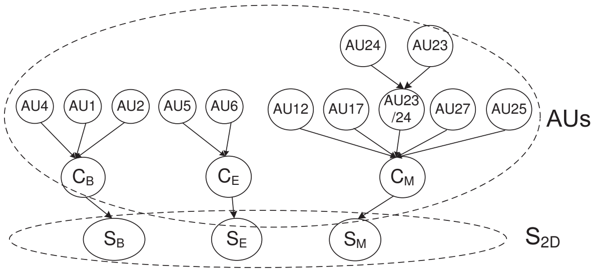


Figure 3. The relationships between the facial muscular movements (AUs) and the 2D local facial component shapes through the intermediate nodes (C_B, C_E, and C_M).

Generally speaking, there are two types of temporal dependencies among AUs: intra-dependency and inter-dependency. Intra-dependency characterizes the self evolution of an AU, while inter-dependency captures temporal dependencies among different AUs, i.e., an AU will be activated following the activation of another AU. For example, in a spontaneous smile, AU6 (cheek raiser) is activated in an average of 0.4 second after the activation of AU12 (lip corner puller) [17]; after both the actions reach their apexes simultaneously, AU6 is relaxed before AU12 is released. Furthermore, due to the variability among individuals and different contexts, the dynamic relationships among AUs are stochastic. Therefore, systemically capturing the spatiotemporal dependencies among AUs and incorporating them into facial action recognition process is especially important for spontaneous facial behaviors.

3.3.1 A DBN for Modeling Semantic and Dynamic Dependencies among AUs

In this paper, we propose to use a DBN to model and learn both the spatial and dynamic dependencies among AUs. A DBN is a directed acyclic graphical model, which models the spatiotemporal dependencies of a set of random variables $\mathbf{X}$ over time [6]. Let $\mathbf{X}^t$ represents 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 Conditional Probability Tables (CPTs) for all nodes. There are two assumptions in the DBN model: first, we assume that the system is first-order Markovian, i.e., $P(\mathbf{X}^{t+1}|\mathbf{X}^0, \cdots, \mathbf{X}^t) = P(\mathbf{X}^{t+1}|\mathbf{X}^t)$; and second, we assume 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 $B_0 = (G_0, \Theta_0)$ captures the static distribution over all variables $\mathbf{X}^0$; and (2) the transition network $B_\rightarrow = (G_\rightarrow, \Theta_\rightarrow)$ specifies $P(\mathbf{X}^{t+1}|\mathbf{X}^t)$ for all t in a finite time slices $T+1$.

Given a DBN model, the joint probability over all variables $\mathbf{X}^0, \cdots, \mathbf{X}^T$ can be factorized as follows:

$$P(\mathbf{x}^0, \cdots, \mathbf{x}^T) = P_{B_0}(\mathbf{x}^0) \prod_{t=0}^{T-1} P_{B_\rightarrow}(\mathbf{x}^{t+1}|\mathbf{x}^t), \qquad (2)$$

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

$$P_{B_\rightarrow}(\mathbf{x}^{t+1}|\mathbf{x}^t) = \prod_{i=1}^{N} P_{B_\rightarrow}(x_i^{t+1}|pa(X_i^{t+1})), \qquad (3)$$

where $pa(X_i^{t+1})$ represents the parent configuration of node 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 j^{th} parent configuration of variable X in a given network.

In this work, B_0 is used to capture the spatial dependencies among AUs, while $B_\rightarrow$ is used to capture and model the intra-dependency and inter-dependency for AUs in two adjacent time slices. The intra-dependency is modeled as an arc linking AU_i 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 inter-dependency is modeled as an arc from AU_i^{t-1} to AU_j^t and represents the pairwise dynamic dependency between two different AUs.

3.3.2 Constructing the Initial DBN

In this work, each AU is represented by a binary value $[0, 1]$ for its presence/absence status. First, we derive an initial static BN to model the semantic AU relationships based on the data analysis from a spontaneous facial expression database in a way similar to [22]. The details of the database are discussed in Section 6. Second, we also need to construct an initial transition network for modeling the dynamic dependencies among AUs.

Since the state 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 is computed as follows:

$$P(AU_j^t = 1|AU_j^{t-1} = 1, AU_i^{t-1} = 1) = \frac{N_{AU_j^t + AU_j^{t-1} + AU_i^{t-1}}}{N_{AU_j^{t-1} + AU_i^{t-1}}}, \qquad (4)$$

where $N_{AU_j^{t-1} + AU_i^{t-1}}$ is the total number of the events that AU_j and AU_i both are 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 t^{th} slice while AU_j and AU_i are present in the $(t-1)^{th}$ slice. The other probabilities are computed similarly.

The initial intra-dependency and inter-dependency among AUs are partially learned from the spontaneous facial expression database. If $P(AU_j^t = 1|AU_j^{t-1} = 0, AU_i^{t-1} = 1)$ is higher than a predefined threshold T_{up} or $P(AU_j^t = 1|AU_j^{t-1} = 1, AU_i^{t-1} = 0)$ 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 a link from AU_i^{t-1} to AU_j^t in the DBN. In this way, an initial DBN is manually constructed as in Figure 4a.

3.3.3 Learning DBN Model

Given a set of observed data D, we can refine the initial DBN model with a structure learning algorithm. 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. Since learning methods for both parts of the model are similar, here we only discuss learning the transition model.

The transition network $B_\rightarrow$ consists of two types of links. Inter-slice links connect the temporal variables of two adjacent time slices. Intra-slice links connect the variables within same time slice, and are same as the static network. To evaluate the fitness of the transition network to the data, it needs to define a scoring function. The score of the $B_\rightarrow$ is defined based on the Bayesian Information Criterion (BIC) [18] as follows:

$$Score_{B_\rightarrow} = logP(B_\rightarrow) + \sum_{i,j,k} N_{\overrightarrow{i},j,k} log\hat{\theta}_{\overrightarrow{i},j,k} - \frac{log(M-S)}{2} K_{B_\rightarrow} \qquad (5)$$

where $logP(B_\rightarrow)$ is the log prior probability of $B_\rightarrow$; the second term evaluates how well $B_\rightarrow$ fits the training data; the third term is a penalty relating to the complexity of the network. $M-S$ is the total number of pair-wise transitions between two adjacent slices in the training data; $K_{B_\rightarrow}$ is the number of parameters in $B_\rightarrow$; $\theta_{\overrightarrow{i},j,k} = P(X_i^t = k|pa^j(X_i^t))$ represents the model parameter for the node X_i^t at k^{th} state given its j^{th} parent configuration in $B_\rightarrow$; and $N_{\overrightarrow{i},j,k}$ accounts for the number of the instances of transition, where X_i^t is at its k^{th} state with its j^{th} parent configuration $pa^j(X_i^t)$, in the training data.

Given the definition of a score for $B_\rightarrow$ as in Eq. (5), we need to identify a structure of $B_\rightarrow$ with the highest score by a searching algorithm subject to some coherent constraints on $B_\rightarrow$. First, the variables $\mathbf{X}^0$ do not have parents. Second, the inter-slice links can only from the previous time slice to current time slice. Finally, both the inter-slice links and intra-slice links should be repeated for $t \in [1, T]$. Further-

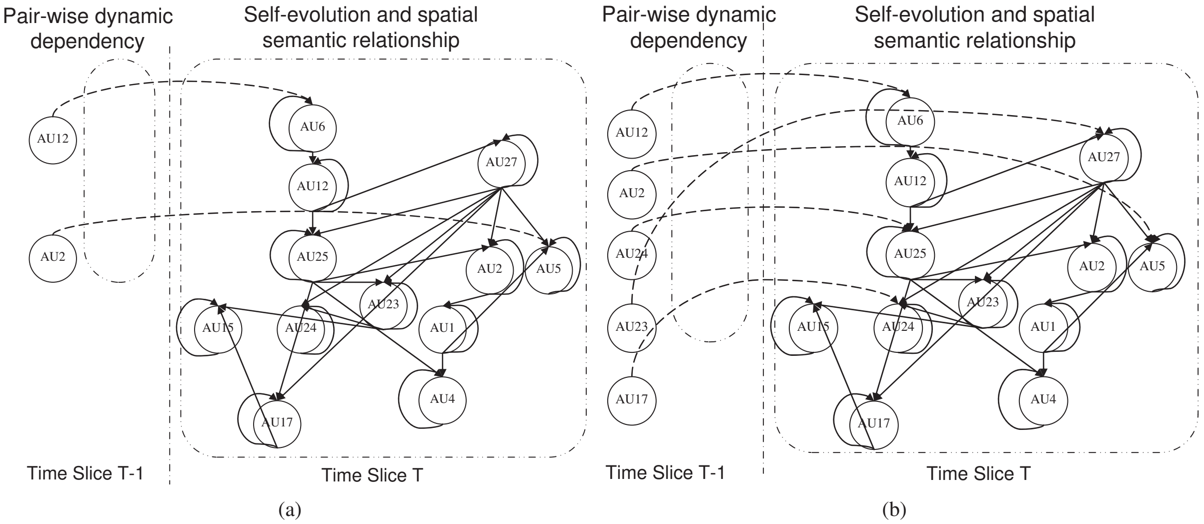


Figure 4. (a) The initial DBN and (b) the learned DBN by the proposed algorithm for AU modeling. The self-arrow at each AU node indicates the intra-dependency. The dashed line with arrow across two time slices indicates the pairwise inter-dependency between different AUs.

more, since we intend to capture the strong dynamic dependencies among AUs that are true for most people, an additional constraint is imposed so that each node X_i^t has at most two parents from the previous time slice. Starting from the initial $B_\rightarrow$ as shown in Figure 4a, we then apply a hill climbing technique [16] to identify the structure of $B_\rightarrow$.

Compared with the manually constructed initial $B_\rightarrow$ in Figure 4a, the learned transition network better reflects the dynamic relationships among AUs in the training data. For example, 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 that the chin boss is already moved upward by activating AU17 (chin raiser). Furthermore, given the scoring function of the static network, the definition of which is similar to that of the transition network, learning the static network is performed similarly as learning the transition network. Figure 4b shows the learned DBN by the proposed learning algorithm.

3.4. Modeling Facial Motion Measurements

In the facial action model, head pose, AUs, 3D facial shape, 2D global facial shape, 2D local facial component shapes, and the intermediate nodes are hidden nodes. To acquire the measurements for the hidden nodes except for the intermediate nodes, we first perform face and eye detection and detect the 28 feature points on neutral face with frontal view. The measurement of 3D facial shape is obtained by personalizing a trained generic 3D shape model. Afterward, the measurements of 2D global shape and local facial component shapes are obtained by tracking the 28 feature points in each frame. Furthermore, based on the personalized 3D facial shape and the tracked global feature points, three face pose angles (i.e., pan, tilt and roll) are estimated by using a technique similar to [4]. The continuous pan angle is further discretized into frontal, left, and right face pose measurement to model the left-right head rotation. Given the normalized face image, we also extract the measurement for each AU based on Gabor wavelet-based feature representation and AdaBoost classification similar to the work in [3]. To incorporate these measurements into the facial action model, a node (shaded) is introduced to represent each measurement. In addition, a link between a hidden node and its measurement is introduced to model the measurement accuracy.

4. A Complete Facial Action Model and Its Parametrization

4.1. A Comprehensive Facial Action Model

Now we have extended and enriched the causal relationships in Figure 1a to a complete DBN model for facial action understanding, as shown in Figure 5. Specifically, the interactions among nonrigid facial muscular movements are characterized by the static links among AUs in the same time slice and the temporal links among AUs across consecutive time slices. The 2D shape deformations of the facial components are controlled by both the head pose through the 2D global shape and the related AUs through the intermediate nodes. In this way, the interactions between head pose (rigid motion) and the AUs (nonrigid motions) are indirectly modeled through their relationships with 2D global and local facial component shapes. Finally, the facial motion measurements are systematically incorporated into the model through the shaded nodes. This model, therefore, completely characterizes the spatiotemporal dependencies between rigid and nonrigid facial motions and accounts for the uncertainties in facial motion measurements.

4.2. Model Learning And Parametrization

Given the model structure shown in Figure 5, we need to define the states for each node and, then, learn the model parameters associated with each node. For each node without parents, it is parameterized by its prior probability. For the continuous node X with discrete/continuous parents, it is characterized by Conditional Probabilistic Distribution (CPD) and defined as $p(X|pa(X))$; whereas for the discrete node with discrete parents, it is characterized by the CPT defined as $p(X|pa(X))$ similar to CPD.

Specifically, head pose is represented by three states: left, frontal, and right in the proposed system. The prior information of the pose $p(Pose)$ can be learned from training data. 3D facial shape S_{3D} is characterized by a continuous 3D shape vector consisting of 28 feature points from neutral faces. 2D global shape $\mathbf{S}_g$ is represented by a continuous shape vector consisting of global feature points, whereas 2D local shape of the j^{th} facial component is represented by a continuous shape vector $\mathbf{S}_{l_j}$ containing the corresponding local feature points. Each AU has two discrete states representing the presence/absence state of the AU. 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 facial component.

Given $Pose = k$ and $\mathbf{S}_{3D} = s_{3D}$, the CPD of $\mathbf{S}_g$ can be defined as follows [12]:

$$p(\mathbf{S}_g = s_g | Pose = k, \mathbf{S}_{3D} = s_{3D}) = (2\pi)^{-\frac{d_g}{2}} |\Sigma_{gk}|^{-\frac{1}{2}} \tag{6}$$
$$exp(-\frac{(s_g - \mathbf{W}_{gk} * s_{3D} - \mu_{gk})^T \Sigma_{gk}^{-1} (s_g - \mathbf{W}_{gk} * s_{3D} - \mu_{gk})}{2})$$

where d_g is the dimension of $\mathbf{S}_g$; μ_{gk}, $\mathbf{W}_{gk}$, and Σ_{gk} are the mean shape vector, regression matrix, and covariance matrix, respectively. Based on the conditional independence in the BN, we can learn μ_{gk}, $\mathbf{W}_{gk}$ and Σ_{gk} locally from the training data consisting of $\mathbf{S}_g$, S_{3D}, and head pose.

For each local shape component node (i.e., $Eyebrow$, Eye, $Nose$, and $Mouth$), its CPD is parameterized as a Gaussian distribution. Given the training data of each 2D local shape, $\mathbf{S}_g$, and the related AUs, we can learn its CPD locally. The CPT $p(C_i|pa(C_i))$ for each intermediate node is manually specified based on the data analysis. The CPD/CPT of each measurement node is learned to reflect the measurement accuracy of the computer vision technique. The CPTs for all the AUs including the static and dynamic links among AUs are learned simultaneously in the local DBN model, as shown in the Figure 4b. Finally, we learn the transition probability $p_{B_\rightarrow}(X_i^{t+1}|pa(X_i^{t+1}))$ for the other temporal links of the DBN.

5. Facial Action Inference

Once the measurement nodes are observed, we can infer the facial action by maximizing the joint probability of pose and AUs given the measurements. 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: $\mathbf{O}_{S_{3D}}$, $\mathbf{O}_{Pose}^{1:t}$, $\mathbf{O}_{S_g}^{1:t}$, $\mathbf{O}_{S_{l_{1 \cdots L}}}^{1:t}$, $\mathbf{O}_{AU_{1 \cdots N}}^{1:t}$ for the measurements of the 3D facial shape, the head pose, the 2D global shape, the 2D local facial component shapes, and the AUs, respectively, where N is the number of target AUs and L is the number of local facial component shapes, the probability $p(Pose^t, AU_{1 \cdots N}^t | \mathbf{O}_{S_{3D}}, \mathbf{O}_{Pose}^{1:t}, \mathbf{O}_{S_g}^{1:t}, \mathbf{O}_{S_{l_{1 \cdots L}}}^{1:t}, \mathbf{O}_{AU_{1 \cdots N}}^{1:t})$ can be factorized and computed via the facial action model by performing the DBN updating process as described in [11]. Then, the true joint states of *Pose* and the AUs are inferred simultaneously over time by maximizing $p(Pose^t, AU_{1 \cdots N}^t | \mathbf{O}_{S_{3D}}, \mathbf{O}_{Pose}^{1:t}, \mathbf{O}_{S_g}^{1:t}, \mathbf{O}_{S_{l_{1 \cdots L}}}^{1:t}, \mathbf{O}_{AU_{1 \cdots N}}^{1:t})$.

Therefore, the true joint states of head pose and the AUs can be inferred simultaneously, given the measurements of the 3D face, head pose, the 2D global shape, the 2D local shapes, and the AUs through probabilistic inference.

6. Experiments on Spontaneous Facial Action Recognition

To demonstrate the system robustness for recognizing spontaneous facial action, the system is trained and tested on a spontaneous facial expression database, which consists of image videos collected through three sources: (1) Multiple Aspects of Discourse research lab at the University of Memphis [1]; (2) Belfast natural facial expression database [7]; and (3) videos obtained from the website. In these image sequences, the subjects are displaying various spontaneous facial expressions with natural head movements. Figure 6 shows an image sequence from the Belfast database [7], where the subject is talking with natural head movement. For this study, all the image sequences 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, while the negative samples are defined as the images without the target AU. For training the 2D facial shapes, we also manually marked the 28 feature points on some images in the database.

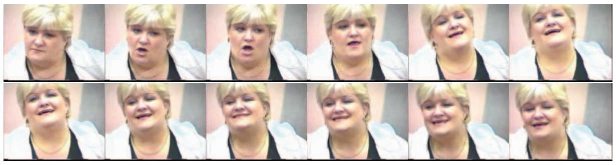


Figure 6. An example image sequence from Belfast database where the subject is talking with natural head movement.

Currently, the database contains 74 image sequences

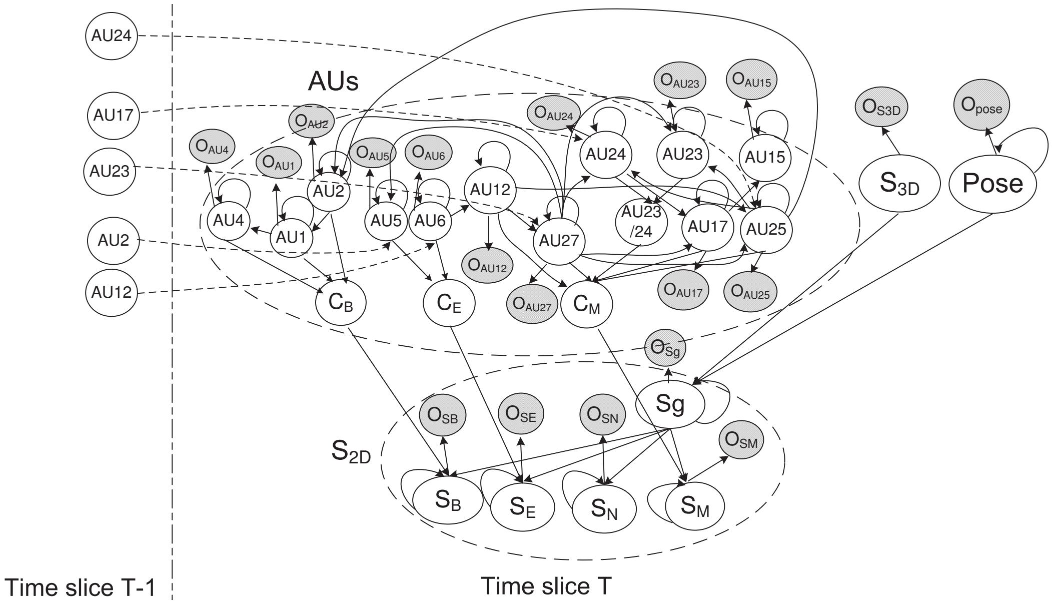
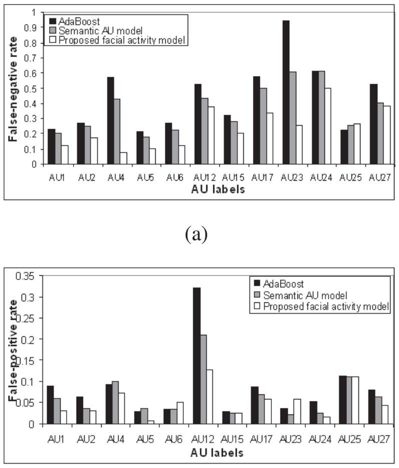


Figure 5. The DBN for spontaneous facial action understanding. The shaded node indicates the observation for the connected hidden node. The self-arrow at the hidden node represents its temporal evolution from time $t-1$ to time t. The link from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the dynamic dependency between different AUs.

from 13 subjects, where 63 image sequences are used for training, and 11 image sequences for testing. In this work, we intend to recognize 12 target AUs, as shown in Figure 5, which frequently occur in the database. Since we intend to recognize AUs under varying head pose, the AdaBoost classifiers are trained on frontal, left, and right face view, respectively, for each AU. Assuming that face pose varies smoothly over time, the AdaBoost classifier corresponding to the face pose estimated in the previous frame is selected to obtain the AU measurement for the current frame.

Figure 7 shows the average AU recognition performance on the spontaneous facial expression database of using the AdaBoost classifiers alone, using the semantic AU model that focuses on modeling the static relationships among AUs and self-development of AUs as in [22], and using the proposed facial action model, respectively. The AdaBoost classifiers achieve an average false negative rate (the ratio of the misclassified positive samples to the total positive samples) of 44% and an average false positive rate (the ratio of the misclassified negative samples to the total negative samples) of 8.58% for the 12 target AUs. By employing the semantic relationships among AUs, the semantic AU model decreases the average false negative rate to 36.5% and decreases the average false positive rate to 6.6%. With the semantic and dynamic relationships among AUs and the interactions between the rigid and nonrigid motions, the average false negative rate further decreases to 24.3%, and the false positive rate decreases to 5.3%.

Compared to the recognition results using the AdaBoost classifiers, the system performance is greatly improved by using the proposed facial action model for some AUs. For example, the false negative rate of AU23 (lip tighten) decreases from 94.4% to 25.9% with a moderate increasing of false positive rate (from 3.6% to 5.8%); the false negative rate of AU12 (lip corner puller) decreases from 53% to 37.8% with a significant drop of false positive rate (from



(a)

(b)

Figure 7. AU recognition results on spontaneous facial expression database: (a) average false negative rates, and (b) average false positive rates. In each figure, the black bar denotes the result using the AdaBoost classifier, the grey bar represents the result using the semantic AU model [22], and the white bar represents the result using the proposed facial action model.

32% to 12.8%); and the false negative rate of AU2 (outer brow raiser) decreases from 26.9% to 16.9% with a decreasing of false positive rate from 6.2% to 3.1%. Furthermore, since the spontaneous facial action is often accompanied by natural head movements, only employing the relationships among AUs are not sufficient to deal with the facial appearance variations due to varying head pose. By incorporating the relationships between head pose and AUs through the 2D facial shapes, the complete facial action model further improves the system performance compared to the recognition results by using the semantic AU model [22]. That demonstrates the effectiveness and importance to model the interactions among the rigid motion and nonrigid motions and to model the spatiotemporal relationships among AUs for understanding spontaneous facial action.

7. Conclusion and Future Work

In this paper, we propose a novel approach for spontaneous facial action analysis and understanding. Specifically, we introduce a probabilistic facial action model based on a DBN to systematically model rigid and nonrigid facial motions, their spatiotemporal interactions, and their image observations. The experiments show that the proposed system yields significant improvements for spontaneous facial action recognition over the method that recognize facial action individually and the method that only focus on one aspect of facial action. The performance improvements come mainly from combining the facial action model with the facial measurements: the erroneous AU measurements can be compensated by the model's built-in spatiotemporal relationships among AUs and the built-in relationships between rigid and nonrigid motions. The important lesson we can learn from this work is that for a robust visual interpretation and understanding, solely improving the computer vision techniques will not be enough. It is important to capture the prior knowledge or context in a probabilistic manner and systematically combine the captured knowledge with the improved visual measurements to achieve a robust and accurate visual interpretation.

References

[1] Multiple aspects of discourse research lab http://madresearchlab.org/.

[2] In P. Ekman and E. Rosenberg, editors, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford Universtity Press, Oxford, UK, 2005.

[3] 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*, 1(6):22–35, September 2006.

[4] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. *Proc. Int'l Conf. on Computer Vision*, pages 323–328, 1998.

[5] 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. SMC*, 1:610–616, 2004.

[6] T. Dean and K. Kanazawa. Probabilistic temporal reasoning. *the Seventh National Conf. on Artificial Intelligence (AAAI88)*, pages 524–528, 1988.

[7] E. Douglas-Cowie, R. Cowie, and M. Schroeder. The description of naturally occurring emotional speech. *Fifteenth Int'l Congress of Phonetic Sciences*, 2003.

[8] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: the Manual*. Research Nexus, Div., Network Information Research Corp., Salt Lake City, UT, 2002.

[9] 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*, 18(4):423–435, 2005.

[10] 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*, pages 195–202, 2003.

[11] K. B. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence*. Chapman and Hall/CRC, London, UK, 2004.

[12] K. Murphy. Inference and learning in hybrid bayesian networks. *Technical Report CSD-98-990, Department of Computer Science, U. C. Berkeley*, 1998.

[13] 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*, pages 206–211, April 1998.

[14] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, 2007.

[15] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[16] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, N.J., 1995.

[17] K. Schmidt and J. Cohn. Dynamics of facial expression: Normative characteristics and individual differences. *IEEE Int'l Conf. on Multimedia and Expo*, pages 728–731, 2001.

[18] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[19] 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*, pages 517–522, 2004.

[20] Y. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In S. Li and A. Jain, editors, *Handbook of face recognition*. Springer, 2004.

[21] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, in Press.

[22] 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*, 29(10):1683–1699, October 2007.

[23] Y. Tong, W. Liao, Z. Xue, and Q. Ji. A unified probabilistic framework for facial activity modeling and understanding. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.

[24] 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. on Multimodal Interfaces*, pages 162–170, 2006.

[25] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. S. Huang. Spontaneous emotional facial expression detection. *J. of Multimedia*, 1(5):1–8, 2006.