A Unified Probabilistic Framework for Facial Activity Modeling and Understanding

Yan Tong Wenhui Liao Zheng Xue Qiang Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA
jiq@rpi.edu

Abstract

Facial activities are the most natural and powerful means of human communication. Spontaneous facial activity is characterized by rigid head movements, non-rigid facial muscular movements, and their interactions. Current research in facial activity analysis is limited to recognizing rigid or non-rigid motion separately, often ignoring their interactions. Furthermore, although some of them analyze the temporal properties of facial features during facial feature extraction, they often recognize the facial activity statically, ignoring the dynamics of the facial activity.

In this paper, we propose to explicitly exploit the prior knowledge about facial activities and systematically combine the prior knowledge with image measurements to achieve an accurate, robust, and consistent facial activity understanding. Specifically, we propose a unified probabilistic framework based on the dynamic Bayesian network (DBN) to simultaneously and coherently represent the rigid and non-rigid facial motions, their interactions, and their image observations, as well as to capture the temporal evolution of the facial activities. Robust computer vision methods are employed to obtain measurements of both rigid and non-rigid facial motions. Finally, facial activity recognition is accomplished through a probabilistic inference by systemically integrating the visual measurements with the facial activity model.

1. Introduction

Face plays an essential role in human communications. In a spontaneous facial behavior, facial activity is characterized by rigid head movement, non-rigid facial muscular movements, and their interactions. Rigid head movement characterizes the overall 3D head pose including rotation and translation. Non-rigid facial muscular movement resulting from the contraction of facial muscles, characterizes the local facial activity in a finer level. Based on the facial action coding system (FACS) [5], the non-rigid facial muscular movement could be described by a set of facial action units (AUs). A system that could automatically infer 3D facial activities from the captured 2D images in real time has applications in a wide range of areas such as automated tools for human behavior science, human-computer interaction, interactive games, computer-based learning, entertainment, and medicine.

However, developing such a system faces several challenges. Firstly, facial activities are rich and complex. For example, there are thousands of distinguished non-rigid facial muscular movements in our daily life, and most of them differ in subtle changes. Secondly, rigid and non-rigid motions are often non-linearly coupled together in the captured 2D images. Thirdly, the visual observations of facial activities are often uncertain and ambiguous. Finally, facial activity evolves over time, and therefore recognizing facial activity requires accounting for its temporal evolution.

Extensive research has been devoted to recognizing the facial activities. Assuming that the face variations caused by rigid and non-rigid facial motions are independent of each other, traditional methods recognize the rigid motion and non-rigid motions sequentially and separately [2, 4, 14, 13, 6], while ignoring the interactions among them. However, the rigid and non-rigid facial motions intertwine with each other, and it is their interaction that creates a coherent and meaningful facial display. By ignoring the interactions among facial motions, the current methods for facial analysis are, therefore, inadequate. In addition, the computer vision measurements of the facial activity are always uncertain and ambiguous. They are uncertain because of the presence of noise, occlusion, the complexity with facial activity, and of the imperfect nature of the vision algorithms. They are ambiguous because they only measure certain aspects of the visual activity. These uncertain and ambiguous measurements would not be effective if used alone. They need to be combined through a model of visual activity to better infer facial activity.

Finally, facial activity evolves over time and it can therefore be better characterized by a spatio-temporal pattern instead of only a spatial pattern. Recognizing a spatio-

1-4244-1180-7/07/$25.00 ©2007 IEEE

temporal pattern requires the integration of evidence over time. Therefore, understanding facial activity requires not only estimating each element individually and statically, but more importantly, characterizing the comprehensive interactions among them, as well as their temporal evolutions. This motivates us to treat facial activity recognition in a global context, explicitly exploit and probabilistically model the context, and perform visual recognition within the context through a probabilistic inference.

We propose a unified probabilistic framework based on the dynamic Bayesian network (DBN) to simultaneously and coherently represent rigid and non-rigid facial motions, their interactions, and their image observations. The framework also captures the temporal evolution of rigid and non-rigid motions as well as the uncertainties with image observations. Finally, facial activity recognition is accomplished through a probabilistic inference by systematically integrating the visual measurements with the facial activity model.

2. Related Work

In general, the previous work on facial activity analysis could be classified into two groups. The traditional methods [2, 4, 14, 13, 6] assume 3D head pose is independent of non-rigid motions, and estimate 3D pose and non-rigid motions separately in two steps: usually a tracking process is performed firstly, and 3D pose is estimated from tracked salient facial feature points; then facial expression is recognized from the pose-free facial texture or from the extracted non-rigid motions by eliminating the effect of pose. However, since rigid motion and non-rigid motions are non-linearly coupled in the projected 2D facial shape/appearance, head pose estimation is not reliable under varying facial expressions. Likewise, facial expression recognition is not accurate, since it is difficult to isolate the motion due to facial expression from the one due to head movement. Therefore, most of the research in this group is limited to either estimating head poses on neutral faces or recognizing facial expressions/facial action units on frontal view faces.

Recently, research has been carried out to explicitly model the coupling between the rigid head movement and non-rigid motions for facial activity analysis. Among them, [3, 15, 1] estimate the rigid motion and non-rigid motions simultaneously based on a 3D face model by modeling the interaction between 3D head pose and facial expression as a non-linear function. Specifically, Bascle and Blake [3] assume facial expression and 3D head pose could be represented by a bilinear equation, and decouple them by singular value decomposition. Zhu and Ji [15] solve the 3D pose and facial expression parameters respectively by a non-linear decomposition method. Similarly, Anisetti et al. [1] propose a 3D expressive face model to account for the person-dependent shape and facial animation units, while human intervention is required in the first frame. Although the above methods successfully decouple the rigid and the non-rigid facial motions, facial expression and head movements are recognized independently from the recovered rigid and non-rigid motions separately. These approaches, therefore, ignore the interaction between the rigid and non-rigid motion as well as the interactions among the non-rigid motions. In addition, these approaches only focus on recognizing the facial activities from facial shape deformations, while the facial appearance variation caused by facial muscular movement (e.g. the wrinkles and bulges) may provide more information on non-rigid facial activity recognition.

Marks et al. [10] model the image sequence as a stochastic process generated by object motion, object texture, and background texture, and track them simultaneously. While they simultaneous recover the 3D head pose and the non-rigid facial deformation, they do not perform facial expression or facial action recognition.

More recently, the research by Tong et al. [12] shows that AU recognition benefits from explicitly modeling the causal relationships among AUs. For example, they show some patterns of AU combinations appear frequently to express meaningful facial expressions while other AUs are rarely present together. By exploiting such co-presence and co-absence relationships, they demonstrate significant improvements in AU recognition, especially for the difficult AUs. However, their research is limited to AU recognition on nearly frontal view faces, ignoring the impact of head movement on AU measurements.

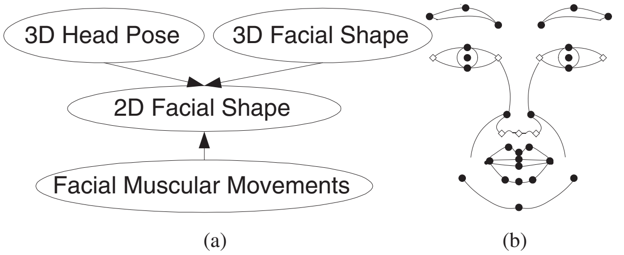
In summary, current work either focuses on recognition of one type of facial motions while ignoring the other one or recognizing both motions separately and ignoring their interactions. Hence, these approaches could not recognize facial activities reliably and robustly. The cornerstone of our system is to treat facial activity recognition in a global context, explicitly exploit and probabilistically model the context, and perform visual recognition within the context through a probabilistic inference. This philosophy deviates significantly from current work in computer vision, which tends to focus narrowly on the target while ignoring the surrounding context.

3. Facial Activity Understanding with A DBN

3.1. Overview of the Facial Activity Model

In the scenario of facial activity analysis, the 2D facial shape could be viewed as a stochastic process generated by three hidden causes: head pose, 3D facial shape, and non-rigid facial muscular movements, which are characterized by a set of action units. Figure 1(a) shows such causal relationships in facial activity.

Given a 3D face, the deformation of a 2D facial shape reflects the action of both head pose and facial muscular movements. Specifically, the head pose and facial muscular movements may affect different sets of facial feature points. Some facial feature points (e.g. the uppermost point on the philtrum and eye corners) are relatively invariant to the facial muscular movements, and their movements are primarily caused by head pose. On the other hand, the others, (e.g. the points on the eye lids and the points on the lips), are not only affected by the head pose, but also sensitive to the facial muscular movements. Based on this observation, the 2D facial shape is represented by a two-level hierarchy including global facial shape and local facial component shapes as shown in Figure 1(b).



(a) (b)

Figure 1. (a) A graphical model to represent the relationships in facial activity. (b) Facial feature points on a frontal view face: the black dot represents the local shape point, while the diamond represents the global shape point.

Hence, the variation of the global shape $\mathbf{S}_g$ characterizes the rigid head movement, while the deformation of the local facial component shapes (i.e. eyes, eyebrows, mouth, nose, and face contour) represents the combined rigid motion and non-rigid facial muscular movements. Based on the semantic relationships shown in Figures 1, we propose to use a Bayesian network (BN) to model 3D facial shape, facial muscular movements, 2D global shape, 2D facial component shapes, and their relationships.

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 conditional dependencies among the rigid motion, non-rigid motions, and their interactions on the 2D facial shapes. Based on the model, facial activity recognition is to find the optimal states of rigid motion (head pose) and non-rigid motions (AUs) by maximizing the joint probability of pose and AUs given their measurements, i.e.

$$pose^*, \mathbf{AU}^* = \underset{pose,\mathbf{AU}}{\operatorname{argmax}}\, p(pose, \mathbf{AU}|O_{pose}, O_{\mathbf{AU}}) \qquad (1)$$

where **AU** is the set of all AUs of interest; O_{pose} and $O_{\mathbf{AU}}$ denote the measurements of the head pose and AUs respectively.

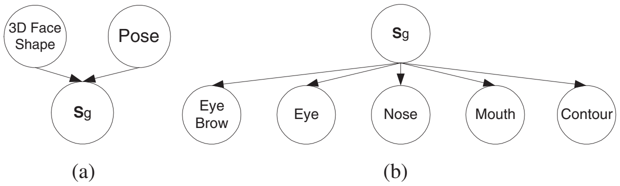
In the next several sections, we gradually show how the relationships in Figure 1(a) can be modeled by a dynamic Bayesian network, based on which we can solve Eq. (1).

3.2. Modeling Rigid Motion with 2D Global Shape

The 2D global shape $\mathbf{S}_g$ is directly affected by the 3D facial shape and the head pose. The 3D facial shape governs the shape of $\mathbf{S}_g$, while the 3D head pose controls both the position and shape of $\mathbf{S}_g$. This causal dependency can be represented by a directed link between the head pose/3D facial shape and the 2D global shape $\mathbf{S}_g$ as shown in Figure 2(a). Given $\mathbf{S}_g$, head pose and 3D facial shape are dependent on each other. Furthermore, the 3D facial shape and the head pose are employed as global constraints for the overall system to produce a globally consistent face shape.

3.3. Modeling the Relationship between 2D Global Shape and Local Facial Components

The local facial components are indirectly affected by the rigid head movement through the 2D global shape $\mathbf{S}_g$. Given the 2D global shape, the position (center) of each local facial component could be roughly estimated. For example, the center of eye could be determined given the eye corners, which are parts of the global shape. Then, this causal relationship can be represented by a link between the global shape and the local facial component as illustrated in Figure 2(b).



(a) (b)

Figure 2. (a) The head pose and 3D facial shape directly affect the 2D global shape $\mathbf{S}_g$. (b) The causal relationship between the global shape $\mathbf{S}_g$ and the local facial component shapes.

3.4. Modeling the Non-rigid Motion with Local Facial Components

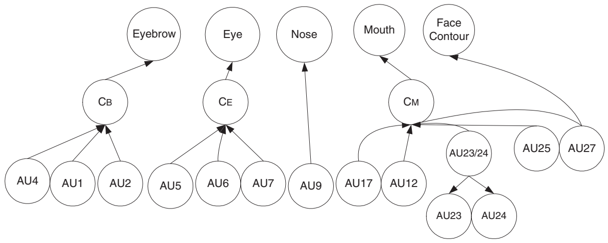


Figure 3. The relationship between the AUs and the local facial component shape. Intermediate nodes (C_B, C_E, and C_M) are introduced to model the correlations among AUs.

Besides the rigid motion, the non-rigid facial muscular movements produce significant changes in the shape of the facial components. These 3D facial muscular movements can be systematically represented by AUs, which are anatomically related to the contraction of the facial muscles as defined in [5]. For example, AU27 (mouth stretch) implies a widely open mouth and the stretched face contour; and AU4 (brow lowerer) makes the eyebrows lower and pushed together. Hence, the shape of each facial component is determined by the related AUs. For instance, there are six AUs controlling the mouth movements, and three AUs for eyebrow movements. In this work, we intend to 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). Details about AUs and their definitions may be found in [5].

We therefore connect the related AUs to the corresponding facial component. For example, AU9 (nose wrinkler) is connected to the nose; while AU1 (Inner brow raiser), AU2 (Outer brow raiser), AU4 (Brow lowerer) are connected to the eyebrow. However, if directly connecting all related AUs to one facial component, too many AU combinations should be considered, while most of them rarely occur in the daily life. For example, based on the analysis of the training data, there are only 8 common AU/AU combinations for the mouth, in spite of 128 potential AU combinations. Thus only a set of common AU combinations, which produce significant non-rigid facial activities, 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 introduced to model the correlations among AUs, and to reduce the number of AU combinations. For example, the intermediate node "C_M" has 8 states, each of which represents a common AU/AU combination controlling mouth movement. Figure 3 shows the modeling of the relationship between the non-rigid facial motions and the local facial component shapes.

3.5. AU Relationships Modeling

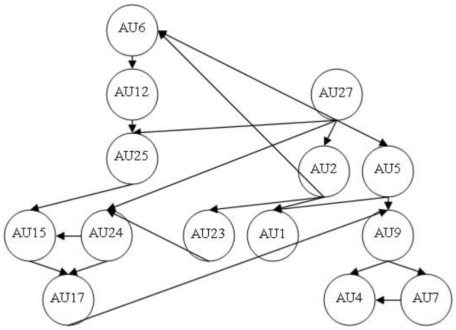


Figure 4. A graphical model for AU relationships modeling (adapted from [12]).

The previous analysis focuses on the relationship between AUs and facial component, but ignoring the relationships among AUs themselves. Based on our previous study in [12], there are semantic relationships among the AUs such as co-presence/absence relationships and mutually exclusive relationships. The co-presence relationships characterize 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 opened, then most likely the eyebrows are raised up, since it implies a surprise expression. On the other hand, some AUs are anatomically mutually exclusive. For instance, the lips cannot be parted as AU25 (lips part) and pressed as AU24 (lip presser) simultaneously. By incorporating such relationships in the proposed model, the interactions among the non-rigid facial motions can be well modeled. In addition, with such relationships, the proposed model could handle the situations where some facial components are occluded by using the information from other facial components. Figure 4 presents a BN for modeling AU relationships.

3.6. Modeling the Dynamics

In the previous discussion, we only focus on modeling the static relationships. However, it lacks the ability to express the temporal dependencies between the consecutive occurrence of the facial activity in an image sequence. Since the facial activity is an event that develops over time, it is better to understand the facial activity from a sequence of observations over time instead of from a snapshot.

Therefore, we extend the static BN model to a dynamic Bayesian network (DBN). In a DBN, the links between the nodes at time $t-1$ and the nodes at time t depict how a random variable at the previous time frame affects the random variable at the current time frame, such that the random variables at time t are not only influenced by the variables at current time frame, but also by the corresponding variables in previous time frame. Such a DBN model is capable of modeling the dynamic aspect of the facial activity.

3.7. Modeling Measurements

In a Bayesian network, the nodes could be grouped into hidden nodes and measurement nodes. The hidden nodes are the head pose, AUs, 3D facial shape, 2D global shape, and 2D local facial component shapes; while the measurement nodes represent their observations obtained through some computer vision techniques respectively. The measurement nodes provide evidence to infer the states of the hidden nodes. The relationships between the measurements and the hidden variables can be established through links, which represent the measurement uncertainty with the computer vision techniques.

We employ various computer vision techniques to acquire various image measurements. Specifically, we first perform face and eyes detection on neutral face with frontal view. Given the detected face and eye centers, we obtain the measurement of the 3D facial shape by personalized a trained generic 3D shape model (face mesh). Given the knowledge of eye centers, the face region is normalized, and is convolved pixel by pixel by a set of multi-scale and multi-orientation Gabor filters. Then the measurements of 2D global shape and local facial components are obtained by detecting/tracking the 34 facial feature points as shown in Figure 1(b) based on Gabor wavelet matching. Active Shape model [7] is also employed as the shape constraint to improve the robustness of facial features tracking. Based on the personalized 3D facial shape and the tracked global facial feature points, three face pose angles (i.e. pan, tilt and roll) are estimated through the weak perspective projection model using a technique similar to [3]. And the continuous pan angle is discretized into frontal, left, and right face pose measurement. Given the normalized face image, we also extract the measurement for each AU through a general purpose learning mechanism based on Gabor feature representation and AdaBoost classifiers similar to [2].

3.8. A Comprehensive Model for Facial Activity Understanding

Now we are ready to present the complete DBN model for facial activity understanding as shown in Figure 5. Specifically, there are three layers in the proposed model: the first layer consisting of the 3D pose, 3D facial shape, and the 2D global shape $\mathbf{S}_g$; the second layer containing a set of 2D local shapes corresponding to the facial components; and the third layer including a set of AUs. We employ the first layer as the global constraint for the overall system, such that it will guarantee globally meaningful facial activity. Meanwhile, the local structural details of the facial components are constrained not only by the local shape parameters, but also by the non-rigid facial muscular movements, represented by the related AUs. In addition, the dependency among different facial components could be represented by the semantic relationships among the AUs.

For presentation clarity, we use self-arrows to indicate the temporal evolution of a temporal variable from the previous time frame to the current time frame. For example, the self-arrow at "$Pose$" means a temporal link connecting "$Pose$" at time $t-1$ to "$Pose$" at time t. Furthermore, we associate each hidden node with a measurement node, which is indicated by a shaded circle in Figure 5.

Such a DBN is capable of systematically and simultaneously representing the relationships among rigid head movement, non-rigid muscular movements, and their interactions on the 2D facial shape, accounting for uncertainties in their measurements, and modeling the dynamic nature of the facial activity.

4. Model Learning And Parameterizing

Given the model structure shown in Figure 5, we need to define the states for each node, and then learn the conditional probability distribution (CPD) associated with each node. The CPD defines conditional probability of each node given its parents $p(X|pa(X))$. Hereafter, $pa(X)$ is defined as the set of parent nodes of node X.

In this work, the head pose is represented by three different views: left, frontal, and right in the proposed system. The prior information of the pose $p(Pose)$ could be learned from the training images. The 3D facial shape S_{3D} is characterized by a generic 3D shape model consisting of 34 facial feature points from neutral faces. The 2D global shape is represented by a shape vector $\mathbf{S}_g$ consisting of global feature points, while the i^{th} local facial component shape is represented by a shape vector $\mathbf{S}_{li}$ containing the corresponding local feature points. And each AU has two states, which represents the presence/absence of the AU.

Given the head pose $pose = k$ and the 3D facial shape $\mathbf{S}_{3D} = s_{3D}$, the CPD of $\mathbf{S}_g$ can be represented as [11]:

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

where d_g is the dimension of the 2D global shape $\mathbf{S}_g$, and γ_{gk}^2 is defined as a Mahalanobis distance:

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

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 could learn the μ_{gk}, $\mathbf{W}_{gk}$ and Σ_{gk} locally as shown in Figure 2(a) given the training data.

The CPT (conditional probabilistic table) $p(C_i|pa(C_i))$ for each intermediate node (i.e. C_B, C_E, and C_M) is manually specified based on the data analysis. For example, we assign $p(C_B = 0|AU1 = 0, AU2 = 0, AU4 = 0) = 1$ for the neutral state of the eyebrow, and $p(C_B = 1|AU1 = 1, AU2 = 1, AU4 = 0) = 1$ for a raised eyebrow.

For the local shape component node ($EyeBrow$, Eye, $Nose$, etc..), its CPD is parameterized as a Gaussian distribution. Specifically, for the $EyeBrow$ node, let $\mathbf{S}_B$ denote the 2D local shape of eyebrow, and assume the CPD of $Eyebrow$ $p(\mathbf{S}_B|\mathbf{S}_g = s_g, C_B = k)$ satisfying a Gaussian distribution with corresponding mean shape vector μ_{bk}, regression matrix $\mathbf{W}_{bk}$, and covariance matrix Σ_{bk}. Then we could learn the parameters μ_{bk}, $\mathbf{W}_{bk}$ and Σ_{bk} locally given the 2D global shape $\mathbf{S}_g = s_g$ and the related AUs. The parameters for Eye, $Nose$, $Mouth$ and $FaceContour$ are defined and learned similarly to those of $Eyebrow$.

The CPTs for all the AUs are learned simultaneously in a local model shown in the Figure 4. Specifically, let θ_{ijk} indicate a probability parameter for an AU node with the graph G as $\theta_{ijk} = p(AU_i = k|pa(AU_i) = j, G)$, where $AU_i = k$ represents the kth state of variable AU_i, and $pa(AU_i) = j$ represents the jth configuration of the parent nodes of AU_i. Thus the goal of learning parameters is to maximize the likelihood $p(D|\theta, G)$ given a database D and the graph G as [8]:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, p(D|\Theta, G) = \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{k=1}^{K} \theta_{ijk}^{C_{ijk}} \quad (4)$$

where N is the number of AUs in the BN; M is the number of the parent instantiations for variable AU_i; K is the number of states of variable AU_i; and C_{ijk} is the number of cases in database D for $AU_i = k$ and $pa(AU_i) = j$.

The CPD of each measurement node given its parent is learned to reflect the measurement accuracy of the computer vision technique. For example, $p(O_{AU_i}|AU_i)$ represents the measurement accuracy with the corresponding AdaBoost classifier. Finally, we learn the transition probability $p(X^t|X^{t-1})$ for each temporal link of the DBN.

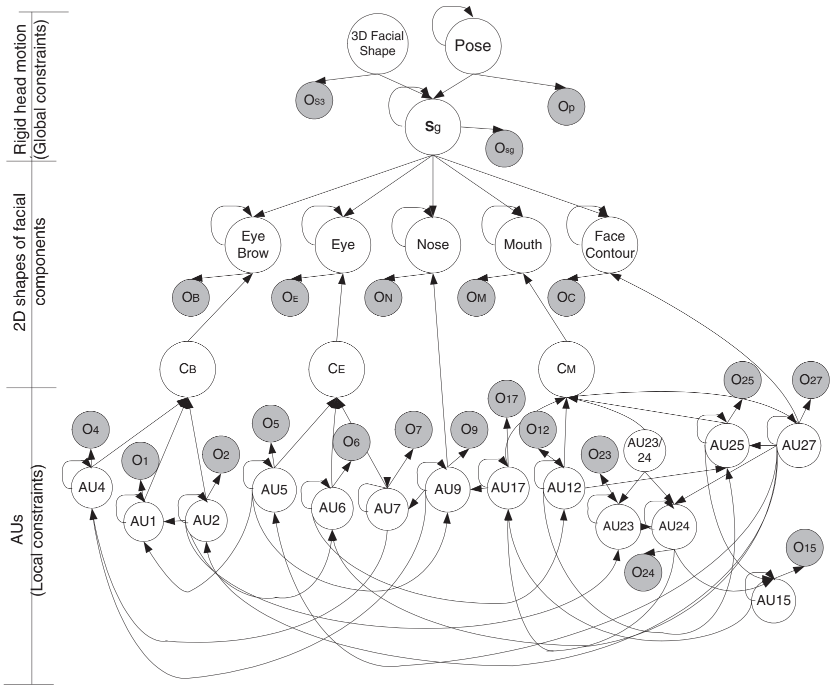


Figure 5. The dynamic Bayesian network for facial activity understanding: the first layer represents the global constraint for the whole system, the second layer represents a set of 2D shapes of facial component, and the third layer characterizes the dependency between the 2D facial component shapes and the non-rigid motions. The shaded node indicates the observation for the connected hidden node. The self-arrow at each hidden node represents its temporal evolution from previous time frame to the current time frame.

5. Facial Activity Inference

Once the measurement nodes are observed, we could infer the facial activity by finding the most probable explanation (MPE) of the evidence as shown in Eq.(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, and thus it finds the most probable state combination of rigid and non-rigid facial motions.

Denote $O_{\mathbf{S}_3}$, O_p, $O_{\mathbf{S}_g}$, $O_{\mathbf{S}_{l_j}}$, and O_{AU_i} as the measurements of the 3D face, head pose, the 2D global shape, the j^{th} 2D local facial component, and AU_i respectively. Based on the conditional independence encoded in the DBN, the inference could be factorized as below:

$$p(pose, AU_{1\cdots N}|O_{\mathbf{S}_3}, O_p, O_{\mathbf{S}_g}, O_{\mathbf{S}_{l_{1\cdots M}}}, O_{AU_{1\cdots N}}) = \tag{5}$$

$$c * \int_{\mathbf{S}_{3D},\mathbf{S}_g,\mathbf{S}_{l_j}} \sum_{C_k} \{p(pose)p(\mathbf{S}_{3D})p(O_{\mathbf{S}_3}|\mathbf{S}_{3D})p(\mathbf{S}_g|\mathbf{S}_{3D}, pose)$$

$$\prod_j^M p(\mathbf{S}_{lj}|pa(\mathbf{S}_{lj})) \prod_k^K p(C_k|pa(C_k)) \prod_i^N p(AU_i|pa(AU_i))$$

$$p(O_{\mathbf{S}_g}|\mathbf{S}_g)p(O_{pose}|pose)[\prod_j^M p(O_{\mathbf{S}_{lj}}|\mathbf{S}_{lj})][\prod_i^N p(O_{AU_i}|AU_i)]\}$$

where c is a normalization constant; M is the number of local facial components; N is the number of target AUs; and K is the number of the intermediate nodes. The factorized probabilities in Eq. (5) are the CPDs that are learned as discussed in Section 4.

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. Experimental Results

6.1. Facial Expression Databases

The proposed system is trained and tested on FACS labeled images from two databases. The first database is Cohn and Kanade's DFAT-504 database [9], which consists of more than 100 subjects. However, the image sequences in Cohn and Kanade's database only contain frontal view face images. In order to demonstrate our system under more natural and realistic circumstance, we also constructed our own database, consisting of 40 image sequences from 8 subjects containing the target AUs. The database is collected under uncontrolled illumination and background. The subjects are instructed to perform the target AUs or the basic facial expressions while turning around their head. Hence, the face undergoes large face pose (−30° to 30° left to right) and significant facial expression changes simultaneously.

In this work, all the image sequences in the two 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 34 feature points on the images from the two databases.

6.2. Evaluation on Cohn and Kanade DataBase

We first evaluate our system on the Cohn-Kanade database [9] 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.

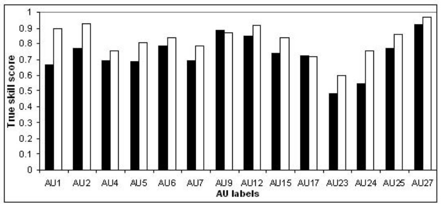


Figure 6. Comparison of AU recognition results on the novel subjects in Cohn-Kanade database using the AdaBoost classifier (black bar), and the proposed model (white bar) based on the true skill score (Hansen Kuiper Discriminant), which is the difference between the positive rate and the false positive rate.

Figure 6 shows the performance for generalization to novel subject in Cohn-Kanade database of using the AdaBoost classifiers alone, and using the proposed model respectively. The AdaBoost classifiers achieve an average positive recognition rate 80.6% and false positive rate 7.84% for the 14 target AUs. With the use of the proposed model, our system achieves an average positive recognition rate 85.4% and false positive rate 3.6%.

6.3. Evaluation under Realistic Environment

In order to demonstrate the robustness of the proposed system, we perform experiments on our own database under more realistic environment, where the face undergoes facial expression and face pose changes simultaneously. The system is evaluated based on the leave-one-subject-out cross-validation. Since we intend to recognize the AUs under varying face pose, the AdaBoost classifiers are trained on frontal, left, and right face view images respectively for each AU. Assuming the face pose varies smoothly over time, the AdaBoost classifier corresponding to the face pose estimated in the previous frame is selected to extract the AU measurement for the current frame.

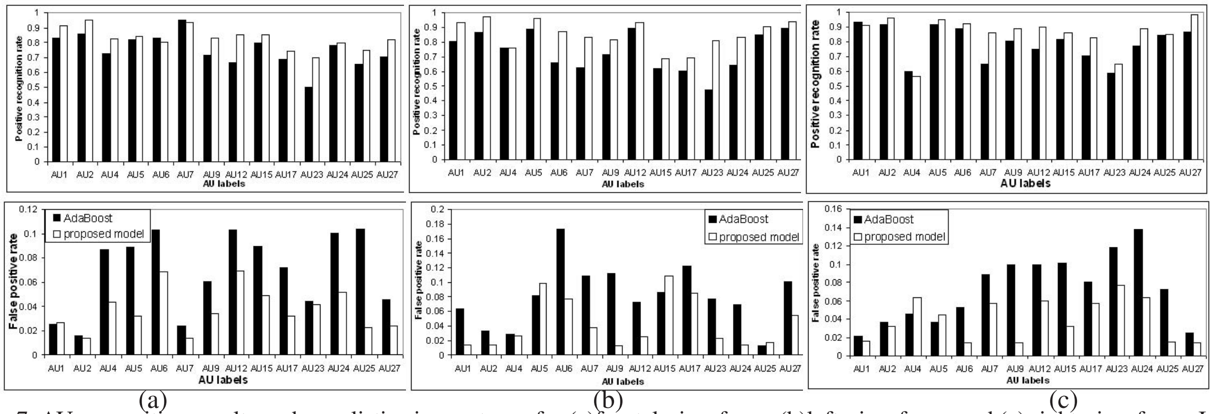
The system performance is reported in Figure 7. Compared to the AU recognition based on AdaBoost only, for the frontal-view face, the proposed model achieves an increase in average positive recognition rate by 7.9% with a decrease in false positive rate by 3.1%; for the right-view face, the proposed method increases the positive rate by 6.9% with a decrease in false positive rate by 3.2%; and for the left-view face, the proposed model significantly improves the positive rate by 11.8% with a decrease in false positive rate by 3.7%. Especially for the AUs that are difficult to be recognized, the system performance is greatly improved. For example, the positive recognition rate of AU23 (lip tighten) is increased from 47.6% to 80.1% with a false positive rate decreasing from 7.7% to 2.3% for the left view face; the positive recognition rate of AU7 (lid tighten) is improved from 62.9% to 83.3% for the right view face; and the positive rate of AU6 (cheek raiser and lid compressor) is increased from 66.0% to 87.5% with a significant drop of false positive rate decreasing from 17.3% to 7.6% for the left view face.

The system enhancement comes mainly from two aspects. Firstly, the erroneous AU measurements could be compensated by the relationships among AUs and the local facial components. For example, it is difficult to recognize AU23 (lip tightener) especially for non-frontal view faces, since the facial appearance changes caused by AU23 are very subtle (e.g. the wrinkles below and above the lips are not noticeable for the non-frontal view faces). However, with the proposed method, its positive recognition rate is significantly improved (47.6% to 80.1% for left-view) by incorporating the information from the local shape deformation due to AU23, with which the mouth appears more narrow than that of without AU23. Secondly, the AU recognition is improved due to the relationships among the AUs. For instance, AU6 (cheek raiser and lid compressor) is hard to be recognized, since there is not sufficient texture information around the cheek and the narrowed eye aperture could also result from AU7 (lid tightener). However, AU6 appears mostly in a happiness expression with AU12 (lip corner puller), which causes significant facial appearance changes and is easier to be recognized. By employing the relationship between AU6 and AU12, the recognition of AU6 improves significantly (positive rate increased from 66.0% to 87.5% for left view).

view	[3]	proposed method
frontal	93%	94.2%
right	94.4%	95.6%
left	86.7%	92.2%

Table 1. Comparison of pose estimation using [3] and the probabilistic inference through the proposed model.

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 similar to [3]. The improvement comes from modeling the interactions of head pose and AUs on the 2D facial shape. The shapes of local facial components are refined by the relationships with AUs. As a result, the erroneous pose measurement is compensated by the improved 2D facial shape. In summary, the proposed system significantly improves AU recognition and pose estimation simultaneously.



(a) (b) (c)

Figure 7. AU recognition results under realistic circumstance for (a)frontal-view faces, (b)left-view faces, and (c) right-view faces. In each figure, the black bar denotes the result by AdaBoost classifier, and the white bar represents the result using the proposed model. The first row demonstrates average positive recognition rate, and the second row displays average false positive rate.

7. Conclusion and Future Work

In this paper, we propose a novel approach for facial activity analysis and understanding. Specifically, we use a dynamic Bayesian network to systematically model the rigid and non-rigid motions, their interactions on 2D facial shapes, the uncertainties with their image observations, as well as their temporal evolution. Under the proposed system, robust computer vision techniques are used to obtain measurements for 3D face, 2D facial shape, head pose, and action units. These measurements are then applied as evidence to the DBN for inferring the rigid and non-rigid facial motions simultaneously. The experiments show the proposed system yields significant improvements in both pose estimation and AU recognition. Currently, we only focus on estimating three different face views. We plan to extend the work to continuous measurement of head pose, which would be more challenging for the DBN modeling procedure and learning approach.

8. Acknowledgement

This paper is supported in part by a grant from ONR under the grant number N00014-03-1-1003.

References

[1] M. Anisetti, V. Bellandi, E. Damiani, and F. Beverina. 3d expressive face model-based tracking algorithm. *Proc. of Signal Processing, Pattern Recognition, and Applications*, pages 111–116, 2006. 2

[2] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*. 1, 2, 5

[3] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. *Proc. of ICCV'98*, pages 323–328, 1998. 2, 5, 7

[4] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition using a particle filter. *Proc. of ICCV'05*, 2:1733–1738, 2005. 1, 2

[5] 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. 1, 3, 4

[6] R. el Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. *Proc. of CVPRW'04*, 2004. 1, 2

[7] T. F.Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models —their training and application. *CVIU*, 61(1):38–59, Jan 1995. 4

[8] D. Heckerman. A tutorial on learning with bayesian networks. *Technical Report MSR-TR-95-06, Microsoft Research*, pages 1–40, 1995. 5

[9] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Proc. of FGR'00*, pages 46–53, 2000. 6, 7

[10] T. K. Marks, J. Hershey, J. C. Roddey, and J. R. Movellan. Joint tracking of pose, expression, and texture using conditionally gaussian filters. *Proc. of NIPS*, 2004. 2

[11] K. Murphy. Inference and learning in hybrid bayesian networks. *Technical Report CSD-98-990, Department of Computer Science, U. C. Berkeley*, 1998. 5

[12] Y. Tong, W. Liao, and Q. Ji. Inferring facial action units with causal relations. *Proc. of CVPR'06*, 2:1623–1630, 2006. 2, 4

[13] 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. of CVPR'05 workshop on Vision for Human-Computer Interaction*, June 2005. 1, 2

[14] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on PAMI*, 27(5):699–714, May 2005. 1, 2

[15] Z. Zhu and Q. Ji. Robust real-time face pose and facial expression recovery. *Proc. of CVPR'06*, 1:681–688, 2006. 2