Simultaneous Facial Feature Tracking and Facial Expression Recognition

Yongqiang Li, Shangfei Wang, Member, IEEE, Yongping Zhao, and Qiang Ji, Senior Member, IEEE

Abstract—The tracking and recognition of facial activities from images or videos have attracted great attention in computer vision field. Facial activities are characterized by three levels. First, in the bottom level, facial feature points around each facial component, i.e., eyebrow, mouth, etc., capture the detailed face shape information. Second, in the middle level, facial action units, defined in the facial action coding system, represent the contraction of a specific set of facial muscles, i.e., lid tighten, eyebrow raiser, etc. Finally, in the top level, six prototypical facial expressions represent the global facial muscle movement and are commonly used to describe the human emotional states. In contrast to the mainstream approaches, which usually only focus on one or two levels of facial activities, and track (or recognize) them separately, this paper introduces a unified probabilistic framework based on the dynamic Bayesian network to simultaneously and coherently represent the facial evolvement in different levels, their interactions and their observations. 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, all three levels of facial activities are simultaneously recognized through a probabilistic inference. Extensive experiments are performed to illustrate the feasibility and effectiveness of the proposed model on all three level facial activities.

Index Terms—Bayesian network, expression recognition, facial action unit recognition, facial feature tracking, simultaneous tracking and recognition.

I. INTRODUCTION

THE recovery of facial activities in image sequence is an important and challenging problem. In recent years, plenty of computer vision techniques have been developed to track or recognize facial activities in three levels. First, in the bottom level, facial feature tracking, which usually detects and tracks prominent facial feature points (i.e., the facial landmarks) surrounding facial components (i.e., mouth, eyebrow, etc.), captures the detailed face shape information. Second, facial actions recognition, i.e., recognize facial Action Units (AUs) defined in the Facial Action Coding System (FACS) [1], try to recognize some meaningful facial activities (i.e., lid tighten, eyebrow raiser, etc.). In the top level, facial expression analysis attempts to recognize facial expressions that represent the human emotional states.

The facial feature tracking, AU recognition and expression recognition represent the facial activities in three levels from local to global, and they are interdependent problems. For example, facial feature tracking can be used in the feature extraction stage in expression/AUs recognition, and expression/AUs recognition results can provide a prior distribution for facial feature points. However, most current methods only track or recognize the facial activities in one or two levels, and track them separately, either ignoring their interactions or limiting the interaction to one way. In addition, the estimates obtained by image-based methods in each level are always uncertain and ambiguous because of noise, occlusion and the imperfect nature of the vision algorithm.

In this paper, in contrast to the mainstream approaches, we build a probabilistic model based on the Dynamic Bayesian Network (DBN) to capture the facial interactions at different levels. Hence, in the proposed model, the flow of information is two-way, not only bottom-up, but also top-down. In particular, not only the facial feature tracking can contribute to the expression/AUs recognition, but also the expression/AU recognition helps to further improve the facial feature tracking performance. Given the proposed model, all three levels of facial activities are recovered simultaneously through a probabilistic inference by systematically combining the measurements from multiple sources at different levels of abstraction.

The proposed facial activity recognition system consists of two main stages: offline facial activity model construction and online facial motion measurement and inference. Specifically, using training data and subjective domain knowledge, the facial activity model is constructed offline. During the online recognition, as shown in Fig. 1, various computer vision techniques are used to track the facial feature points, and to get the measurements of facial motions, i.e., AUs. These measurements are then used as evidence to infer the true states of the three level facial activities simultaneously.

The paper is divided as follows: In Sec. II, we present a brief review on the related works on facial activity analysis; Sec. III describes the details of facial activity modeling, i.e., modeling the relationships between facial features and AUs (Sec. III-B), modeling the semantic relationships among AUs (Sec. III-C), and modeling the relationships between AUs and expressions (Sec. III-D); In Sec. IV, we construct the dynamic dependency and present a complete facial action model; Sec. V shows the
II. RELATED WORKS

In this section, we are going to introduce the related works on facial feature tracking, expression/AUs recognition and simultaneous facial activity tracking/ recognition, respectively.

A. Facial Feature Tracking

Facial feature points encode critical information about face shape and face shape deformation. Accurate location and tracking of facial feature points are important in the applications such as animation, computer graphics, etc. Generally, the facial feature points tracking technologies could be classified into two categories: model free and model-based tracking algorithms. Model free approaches [47]–[49] are general purpose point trackers without the prior knowledge of the object. Each feature point is usually detected and tracked individually by performing a local search for the best matching position. However, the model free methods are susceptible to the inevitable tracking errors due to the aperture problem, noise, and occlusion. Model based methods, such as Active Shape Model (ASM) [3], Active Appearance Model (AAM) [4], Direct Appearance Model (DAM) [5], etc., on the other hand, focus on explicitly modeling the shape of the objects. The ASM proposed by Cootes et al. [3], is a popular statistical model-based approach to represent deformable objects, where shapes are represented by a set of feature points. Feature points are first searched individually, and then Principal Component Analysis (PCA) is applied to analyze the models of shape variation so that the object shape can only deform in specific ways found in the training data. Robust parameter estimation and Gabor wavelets have also been employed in ASM to improve the robustness and accuracy of feature point search [6], [7]. The AAM [4] and DAM [5] are subsequently proposed to combine constraints of both shape variation and texture variation.

In the conventional statistical models, e.g. ASM, the feature points positions are updated (or projected) simultaneously, which indicates that the interactions within feature points are interdependent. Intuitively, human faces have a sophisticated structure, and a simple parallel mechanism may not be adequate to describe the interactions among facial feature points. For example, whether the eye is open or closed will not affect the localization of mouth or nose. Tong et al. [8] developed an ASM based two-level hierarchical face shape model, in which they used multi-state ASM model for each face component to capture the local structural details. For example, for mouth, they used three ASMs to represent the three states of mouth, i.e., widely open, open and closed. However, the discrete states still cannot describe the details of each facial component movement, i.e., only three discrete states are not sufficient to describe all mouth movements. At the same time, facial action units inherently characterize face component movements, therefore, involving AUs information during facial feature points tracking may help further improve the tracking performance.

B. Expression/AUs Recognition

Facial expression recognition systems usually try to recognize either six expressions or the AUs. Over the past decades, there has been extensive research on facial expression analysis [9], [14], [16], [21], [24]. Current methods in this area can be grouped into two categories: image-based methods and model-based methods.

Image-based approaches, which focus on recognizing facial actions by observing the representative facial appearance changes, usually try to classify expression or AUs independently and statically. This kind of method usually consists of two key stages. First, various facial features, such as optical flow [9], [10], explicit feature measurement (e.g., length of wrinkles and degree of eye opening) [16], Haar features [11], [37], Local Binary Patterns (LBP) features [31], [32], Independent component analysis (ICA) [12], feature points [47], Gabor wavelets [14], etc., are extracted to represent the facial gestures or facial movements. Given the extracted facial features, the expression/AUs are identified by recognition engines, such as Neural Networks [15], [16], Support Vector Machines (SVM) [14], [20], rule-based approach [21], AdaBoost classifiers, Sparse Representation (SR) classifiers [33], [34], etc. A survey about expression recognition can be found in [22].

The common weakness of image-based methods for AU recognition is that they tend to recognize each AU or certain AU combinations individually and statically directly from the image data, ignoring the semantic and dynamic relationships among AUs, although some of them analyze the temporal properties of facial features, e.g., [17], [45]. Model-based
methods overcome this weakness by making use of the relationships among AUs, and recognize the AUs simultaneously. Lien et al. [23] employed a set of Hidden Markov Models (HMMs) to represent the facial actions evolution in time. The classification is performed by choosing the AU or AU combination that maximizes the likelihood of the extracted facial features generated by the associated HMM. Valstar et al. [18] used a combination of SVMs and HMMs, and outperformed the SVM method for almost every AU by modeling the temporal evolution of facial actions. Both methods exploit the temporal dependencies among AUs. However, in [26], they only modeled six expressions and they ignored the semantic relationships among AUs, and achieved significant improvement over the image-based method. In this paper, besides modeling the spatial and temporal relationships among AUs, we also make use of the information of expression and facial feature points, and more importantly, the coupling and interactions among them.

C. Simultaneous Facial Activity Tracking/Recognition
The idea of combining tracking with recognition has been attempted before, such as simultaneous facial feature tracking and expression recognition [47], [50], [51], and integrating face tracking with video coding [27]. However, in most of these works, the interaction between facial feature tracking and facial expression recognition is one-way, i.e., facial feature tracking results are fed to facial expression recognition [47], [51]. There is no feedback from the recognition results to facial feature tracking. Most recently, Dornaika et al. [26] and Chen & Ji [30] improved the facial feature tracking performance by involving the facial expression recognition results. However, in [26], they only modeled six expressions and they need to retrain the model for a new subject, while in [30], they represented all upper facial action units in one vector node and in such a way, they ignored the semantic relationships among AUs, which is a key point to improve the AU recognition accuracy.

Compared to the previous related works, this paper has the following features.

1) First, we build a DBN model to explicitly model the two-way interactions between different levels of facial activities. In this way, not only the expression and AUs recognition can benefit from the facial feature tracking results, but also the expression recognition can help improve the facial feature tracking performance.

2) Second, we recognize all three levels of facial activities simultaneously. Given the facial action model and image observations, all three levels of facial activities are estimated simultaneously through a probabilistic inference by systematically integrating visual measurements with the proposed model.

III. FACIAL ACTIVITY MODELING
A. Overview of the Facial Activity Model

1) Single Dynamic Model: The graphical representation of the traditional tracking algorithm, i.e., Kalman Filter, is shown in Fig. 2(a). $X_t$ is the current hidden state, e.g., image coordinates of the facial feature points, we want to track, and $M_t$ is the current image measurement (Hereafter, the shaded nodes represent measurements, i.e., estimates, and the unshaded nodes denote the hidden states). The directed links are quantified by the conditional probabilities, e.g., the link from $X_t$ to $M_t$ is captured by the likelihood $P(M_t|X_t)$, and the link from $X_{t-1}$ to $X_t$ by the first order dynamic $P(X_t|X_{t-1})$.

For online tracking, we want to estimate the posterior probability based on the previous posterior probability and the current measurement

$$P(X_t|M_{1:t}) \propto P(M_t|X_t) \int_{X_{t-1}} P(X_t|X_{t-1})P(X_{t-1}|M_{1:t-1})$$

(1)

$M_{1:t}$ is the measurement sequence from frame 1 to $t$. If both $X_t$ and $M_t$ are continuous and all the conditional probabilities are linear Gaussian, this model is a Linear Dynamic System (LDS).

2) Dynamic Model With Switching Node: The above tracking model has only one single dynamic $P(X_t|X_{t-1})$, and this dynamic is fixed for the whole sequence. But for many applications, we hope that the dynamic can “switch” according to different states. Therefore, researchers introduce a switch node to control the underlying dynamic system [28], [29]. For the switching dynamic model, the switch node represents different states and for each state, there are particular predominant movement patterns. The works in [26] and [30] also
involved multi-dynamics, and their idea can be interpreted as the graphical model in Fig. 2(b). The \( S_t \) is the switch node, and for each state of \( S_t \), there is a specific transition parameter \( P(X_t | X_{t-1}, S_t) \) to model the dynamic between \( X_t \) and \( X_{t-1} \). Through this model, \( X_t \) and \( S_t \) can be tracked simultaneously, and their posterior probability is

\[
P(X_t, S_t | M_{1:t}) \propto P(M_t | X_t) \int_{X_{t-1}, S_{t-1}} P(X_t | X_{t-1}, S_t) \ P(S_t | S_{t-1}) P(X_{t-1}, S_{t-1} | M_{1:t-1})
\]

(2)

In [26], they proposed to use particle filtering to estimate this posterior probability.

3) Our Facial Activity Model: Dynamic Bayesian network is a directed graphical model, and compared to the dynamic models above, DBN is more general to capture complex relationships among variables. We propose to employ DBN to model the spatiotemporal dependencies among all three levels of facial activities (facial feature points, AUs and expression) as shown in Fig. 2(c) [Fig. 2(c) is not the final DBN model, but a graphical representation of the causal relationships between different levels of facial activities]. The \( E_t \) node in the top level represents the current expression; \( AUt \) represents a set of AUs; \( X_t \) denotes the facial feature points we are going to track; \( MAUt \) and \( MXt \) are the corresponding measurements of AUs and the facial feature points, respectively. The three levels are organized hierarchically in a causal manner such that the level above is the cause while the level below is the effect. Specifically, the global facial expression is the main cause to produce certain AU configurations, which in turn causes local muscle movements, and hence feature points movements. For example, a global facial expression (e.g., Happiness) dictates the \( AU \) configurations, which in turn dictates the facial muscle movement and hence the facial feature point positions.

For the facial expression in the top level, we will focus on recognizing six basic facial expressions, i.e., happiness, surprise, sadness, fear, disgust and anger. Though psychologists agree presently that there are ten basic emotions [54], most current research in facial expression recognition mainly focuses on six major emotions, partially because they are the most basic, and culturally and ethnically independent expressions and partially because most current facial expression databases provide the six emotion labels. Given the measurement sequences, all three level facial activities are estimated simultaneously through a probabilistic inference via DBN (section, IV-C). And the optimal states are tracked by maximizing this posterior

\[
E_t^*, AUt^*, X_t^* = \text{argmax}_{E_t, AUt, X_t} P(E_t, AUt, X_t | MAUt, MXt).
\]

(3)

B. Modeling the Relationships Between Facial Features and AUs

In this paper, we will track 26 facial feature points as shown in Fig. 3 and recognize 15 AUs, i.e., AU1, 2, 4, 5, 6, 7, 9, 12, 15, 17, 23, 24, 25, 26 and 27 as summarized in Table I. The selection of AUs to be recognized is mainly based on the AUs occurrence frequency, their importance to characterize the 6 expressions, and the amount of annotation available. The 15 AUs we propose to recognize are all most commonly occurring AUs, and they are primary and crucial to describe the six basic expressions. They are also widely annotated. Though we only investigate 15 AUs in this paper, the proposed framework is not restricted to recognizing these AUs, given an adequate training data set. Facial action units control the movement of face components and therefore, control the movement of facial feature points. For instance, activating AU27 (mouth stretch) results in a widely open mouth; and activating AU4 (brow lowerer) makes the eyebrows lower and pushed together. At the same time, the deformation of facial feature points reflects the action of AUs. Therefore, we could directly connect the related AUs to the corresponding feature points around each facial component to represent the casual relationships between them. Take Mouth for example, we use a continuous node \( X_{\text{Mouth}} \) to represent 8 facial feature points around mouth, and link AUs that control mouth movement to this node. 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, there are eight AUs controlling mouth movement and they collectively produce \( 2^8 \) potential AU combinations. But through the analysis of the database, there
are only eight common AUs or AU combinations for the mouth. Thus, only a set of common AUs or AU combinations, which produce significant facial actions, are sufficient to control the face component movement. As a result, we introduce an intermediate node, e.g., “C_M,” to model the correlations among AUs and to reduce the number of AU combinations. Fig. 4 shows the modeling for the relationships between facial feature points and AUs for each facial component.

Each AU node has two discrete states which represent the “presence/absence” states of the AU. The intermediate nodes (i.e., “C_B,” “C_E,” “C_N,” and “C_M”) are discrete nodes, each mode of which represents a specific AU/AU combination related to the face components. The Conditional Probability Table (CPT) $P(C_i|pa(C_i))$ for each intermediate node $C_i$ is set manually based on the data analysis, where $pa(C_i)$ represents the parents of node $C_i$. For instance, “C_B” has five modes, each of which represents the presence of an AU or AU combination related to the eyebrow movement. We assign the parameter $P(C_B = 0|AU1 = 0, AU2 = 0, AU4 = 0) = 0.9$ to represent the eyebrow at the neutral state, whereas $P(C_B = 1|AU1 = 1, AU2 = 1, AU4 = 0) = 0.9$ to represent that the eyebrow is entirely raised up.

The facial feature nodes (i.e., $X_{Eyebrow}$, $X_{Eye}$, $X_{Nose}$ and $X_{Mouth}$) have continuous states and are represented by continuous vectors, which are the relative image coordinates between the current frame and the neutral frame. Given the local AUs, the Conditional Probability Distribution (CPD) of the facial feature points can be represented as a Gaussian distribution, e.g., for mouth

$$P(X_{Mouth}|C_M = k) \sim N(X_{Mouth}; \mu_k, \Sigma_k)$$ (4)

with the mean shape vector $\mu_k$ and covariance matrix $\Sigma_k$.

The facial feature measurement nodes are continuous vector nodes that have the same dimension as their parents. The CPD for the measurement are modeled as linear Gaussian, e.g., for mouth

$$P(M_{Mouth}|X_{Mouth} = x) \sim N(M_{Mouth}; W \cdot x + \mu_x, \Sigma_x)$$ (5)

with the mean shape vector $\mu_x$, regression matrix $W$, and covariance matrix $\Sigma_x$. These parameters can be learned from training data using expectation maximization (EM) estimation.

C. Modeling Semantic Relationships Among AUs

In the above section, we modeled the relationships between facial feature points and AUs. Detecting each AU statically and individually is difficult due to the variety, ambiguity, and dynamic nature of facial actions, as well as the image uncertainty and individual differences. Moreover, when AUs occur in a combination, they may be nonadditive: that is, the appearance of an AU in a combination is different from its standalone appearance. Fortunately, there are some inherent relationships among AUs, as described in the FACS manual [1]. We can summarize the relationships among AUs into two categories, i.e., co-occurrence relationships and mutual exclusion relationships. The co-occurrence relationships characterize some groups of AUs, which usually appear together to show meaningful facial displays, e.g., AU1+AU2+AU5+AU26+AU27 to show surprise expression; AU6+AU12+AU25 to show happiness expression; AU1+AU4+AU15+AU17 to show sadness 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” [1]. For instance, one can not perform AU25 (lip part) with AU23 (lip tightener) or AU24 (lip pressor) simultaneously. The rules provided in [1] are basic, generic and deterministic. They are not sufficient enough to characterize all the dependencies among AUs, in particular some relationships that are expression and database dependent. Hence, in this paper, we propose to learn from the data to capture additional relationships among AUs.

Tong et al. [25] employed a Bayesian network to model the co-occurrence and mutual exclusion relationships among AUs, and achieved significant improvement for AU recognition compared to image-based methods. Following the work in [25], we also employ a Bayesian network (BN) to model the dependencies among AUs. A BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables. In a BN, its structure captures the dependency among variables, i.e., the dependency among AUs in this paper, and the dependency is characterized by a conditional probability table (CPT), i.e., $\theta$, for each AU node given its parents. Hence, we employ a structure learning algorithm to identify a structure of the DAG, given the training data. The structure learning is to find a structure $G$ that maximizes a score function. In this paper, we employ the Bayesian Information Criterion (BIC) score function [40] which is defined as follows:

$$sD(G) = \max_{\theta} \log P(D|G, \theta) - \frac{\log Num}{2} \cdot Dim_G$$ (6)

where the first term evaluates how well the network fits the data $D$; the second term is a penalty relating to the complexity of the network; $\log P(D|G, \theta)$ is the log-likelihood function of parameters $\theta$ with respect to data $D$ and structure $G$; Num is the number of training data; and $Dim_G$ is the number
of independent parameters, i.e., the number of independent entries in $\theta$.

Cassio et al. [13] developed a Bayesian Network structure learning algorithm which is not dependent on the initial structure and guarantees a global optimality with respect to BIC score. In this paper, we employ the structure learning method [13] to learn the dependencies among AUs. To simplify the model, we use the constraints that each AU node has at most two parents. The learned structure is shown in Fig. 5.

**D. Modeling the Relationships Between AUs and Expression**

In this section, we will add Expression node at the top level of the model. Expression represents the global face movement and it is generally believed that the six basic expressions (happiness, sadness, anger, disgust, fear and surprise) can be described linguistically using culture and ethnically independent AUs, e.g., activating AU6+AU12+AU25 produces happiness expression, as shown in Fig. 6(a).

We group AUs according to different expressions as listed in Table II. But inferring expression from AUs is not simply to transfer the combination of several AUs directly to certain expression. Naturally, combining AUs belonging to the same category increases the degree of belief in classifying to that category, as shown in Fig. 6(a) (the combination of AU6 and AU12 increases the likelihood of classifying as happiness). However, combining AUs across different categories may result in the following situations: First, an AU combination belonging to a different facial expression, e.g., when AU1 occurs alone, it indicates a sadness, and when AU5 occurs alone, it indicates a surprise, however, the combination of AU1 and AU5 increases the probability of fear as shown in Fig. 6(b); Second, increasing ambiguity, e.g., when AU26 (jaw drop), an AU for surprise, combines with AU1, an AU for sadness, the degree of belief in surprise is reduced and the ambiguity of classification may be increased as illustrated in Fig. 6(c).

These relationships and uncertainties are systematically represented by our final facial activity model as shown in Fig. 8. At the top level of the final model, we introduce six expression nodes, (i.e., Surp, Sad, Ang, Hap, Dis and Fea), which have binary states to represent “absence/presence” of each expression. We link each expression node to the corresponding AUs as listed in Table II. The parameter of each expression node is the prior distribution, i.e., $P(Exp)$, and the self dynamic dependency, i.e., $P(Exp_t|Exp_{t-1})$. Expressions are inferred from their relationships with AUs and reasoning over time. In principle, our approach allows a facial expression to be a probabilistic combination of any relevant facial AUs.

**IV. MODELING THE DYNAMIC RELATIONSHIPS**

**A. Constructing Dynamic Structure**

So far, we have constructed a Bayesian network to represent the static relationships among facial feature points, AUs and expressions. In this section, we extend it to a dynamic Bayesian network by adding dynamic links.

In general, a DBN is made up of interconnected time slices of static BNs, and the relationships between two neighboring time slices are modeled by an HMM such that variables at time $t$ are influenced by other variables at time $t$, as well as by the corresponding random variables at time $t - 1$ only. The exact time difference between $t - 1$ and $t$ is determined by the temporal resolution of the image sequence, i.e., the frame rate of the recorded videos, which is critical for setting the temporal relationships. For instance, 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 [21]. Hence, if we choose a small time duration, e.g., a single frame, we may capture many irrelevant events, whereas if we choose many frames as a
duration, the dynamic relationships may not be captured. For instance, Fig. 7 shows how a smile is developed in an image sequence: first, AU12 is contracted at the 4th frame to express a slight smile, and then, AU6 and AU25 are triggered at the 5th and 6th frame respectively to enhance the happiness. As the intensity of happiness increases, AU12 first reaches its highest intensity level, and then, AU6 and AU25 reach their apexes, respectively. Based on this understanding, as well as the temporal characteristics of the AUs we intend to recognize, we empirically set the time duration as 1/6 second.

In the proposed framework, we consider two types of conditional dependencies for variables at two adjacent time slices. The first type, e.g., an arc from AU\textsubscript{i} node at time \( t-1 \) to that node at time \( t \), depicts how a single variable develops over time. For the expression and the facial feature nodes, we only consider this type dynamic. The second type, e.g., an arc from AU\textsubscript{i} at time \( t-1 \) to AU\textsubscript{j} (\( j \neq i \)) at time \( t \), depicts how AU\textsubscript{i} at the previous time step affects AU\textsubscript{j} (\( j \neq i \)) at the current time step. We consider this type dynamic for AU nodes.

The dynamic dependencies among AUs are especially important for understanding spontaneous expression. For example, K. Schmidt et al. [35] found that certain action units usually closely follow the appearance of AU12 in smile expression. For 88% of the smile data they collect, the appearance of AU12 was either simultaneously with or closely followed by one or more associated action units, and for these smiles with multiple action units, AU6 was the first action unit to follow AU12 in 47%. Similar findings are found by Tong et al [20]. Based on this understanding and the analysis of the database, we link AU\textsubscript{2} and AU\textsubscript{12} at time \( t-1 \) to AU\textsubscript{5} and AU\textsubscript{6} at time \( t \) respectively to capture the second dynamic dependencies. Fig. 8 gives the whole picture of the dynamic BN, including the shaded visual measurement nodes. For presentation clarity, we use the self-arrows to indicate the first type of temporal links as described above.

**B. DBN Parameters Learning**

Given the DBN structure and the definition of the CPDs, we need to learn the parameters from training data. In this learning process, we manually labeled the expressions, AUs and facial feature points for some sequences collected from the extended
Cohn and Kanade database [46] frame by frame. Learning the parameters in a DBN is actually similar to learning the parameters for a static BN. During DBN learning, we treat the DBN as an expanded BN consisting of two-slice static BNs connected through the temporal variables, as shown in Fig. 8. Based on the conditional independencies encoded in DBN, we can learn the parameters individually for each local structure. In this way, the quantity of training data required is much smaller than that for a larger network structure. For instance, for the AU and expression model, since all nodes are discrete and let \( \theta_{ijk} \) represent the conditional probability of node \( i \) being in \( k \)th state, given the \( j \)th configuration of its parents
\[
\theta_{ijk} = P(x_i^k|pa^j(X_i))
\]
where \( i \) ranges over all the variables (nodes in the BN), \( j \) ranges over all the possible parent instantiations for variable \( X_i \), and \( k \) ranges over all the instantiations for \( X_i \) itself. Therefore, \( x_i^k \) represents the \( k \)th state of variable \( X_i \), and \( pa^j(X_i) \) is the \( j \)th configuration of the parent nodes of \( X_i \). For example, a node \( AU_{15}^t \) as shown in Fig. 8, represents the presence/absence of AU15 at time step \( t \), with two binary instantiations \((0, 1)\). The parents of \( AU_{15}^t \) are \( AU_7^t, AU_{17}^t, SAD_t \) and \( AU_{15}^{t-1} \), each of which also has two binary instantiations. Hence, there are 16 parent configurations for \( AU_{15}^t \) node.

Given the dataset, the goal of learning parameters is to find the most probable values for \( \theta \). These values best explain the dataset \( D \), which can be quantified by the log likelihood function \( \log(p(D|\theta)) \), denoted as \( L_D(\theta) \). In a BN, every variable is conditionally independent of its non-descendants given its parents (Markov condition), which can be expressed \( P(X_1, \ldots, X_n) = \prod_i P(X_i|pa(X_i)) \) \( (X_i \) represents a variable in the AU and Expression BN network). Based on this property, and assuming that samples are drawn independently from the underlying distribution, we have
\[
L_D(\theta) = \log \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}}
\]
where \( n_{ijk} \) is the count for the case that node \( X_i \) has the state \( k \), with the state configuration \( j \) for its parent nodes; \( n \) is the number of variables (nodes) in the BN; \( q_i \) is the number of parent configurations of \( X_i \) node; \( r_i \) is the number of instantiations of \( X_i \). Since we have complete training data, the learning process can be described as a constrained optimization problem as follows:
\[
\arg\max_\theta L_D(\theta) \quad s.t. \quad g_{ij}(\theta) = \sum_{k=1}^{r_i} \theta_{ijk} - 1 = 0
\]
where \( g_{ij}(\theta) = 0 \) imposes that distributions defined for each variable given a parent configuration sums one over all variable states. This problem has its global optimum solution at \( \theta_{ijk} = n_{ijk}/\sum_k n_{ijk} \).

For the facial feature model, e.g., the Mouth model, we need to learn a mean shape vector and a covariance matrix for each state of the intermediate node, e.g., the \( CM \) node. Since the intermediate node is hidden, in this paper, we employ expectation maximization (EM) estimation to learn these Gaussian parameters. To evaluate the quantity of training data needed for learning the facial activity model, we perform a sensitivity study of model learning on different amounts of training data. For this purpose, the Kullback–Leibler (KL) divergences of the parameters are computed versus the number of training samples. The convergence behaviors for local models, i.e., AUs model, “Eyebrow” model, “Eye” model, “Nose” model, and “Mouth” model, are shown in Fig. 9.

In Fig. 9 we can observe that, when the amount of training data is larger than 3000, all local models converge and have similar K-L divergences. To demonstrate the learning effect, we draw 200 samples from the learned CPDs of the “Mouth” node: \( P(X_{Mouth}|CM) \) as shown in Fig. 10 (The \( X_{Mouth} \) node in our model represents the relative image coordinates. For clarity, we draw the samples by adding a constant neutral shape: \( P(X_{Mouth} + C|CM) \), where \( C \) is a constant neutral shape). From Fig. 10 we can observe that AUs can provide prior distribution for facial feature points, since given different AUs, facial feature point samples drawn from the learnt distribution can reflect the mouth movement shape.

C. DBN Inference

In the above sections, we have learned the DBN model to represent the three level facial activities. During tracking and recognition, this prior DBN model is combined with the likelihood of the measurements, i.e., estimates for each node, to infer the posterior probability. Therefore, the estimation contains two steps in our framework. First, we employ various image-based methods to acquire the necessary estimates. For AUs, we employ a technique based on the AdaBoost classifier
and Gabor features [43] to obtain AU estimates. For facial feature points, we first use the detection method [8] to obtain the feature points on the neutral face (the subject is asked to perform neutral face in the first frame of the sequence). Then the feature points are tracked using the state-of-the-art facial feature tracker [8], which is based on Gabor wavelet matching and active shape model. In this paper, we infer expressions directly from the corresponding AUs, which means we do not employ any image-based method to obtain the estimates for expression nodes. Without node estimates, the hidden expression nodes can still help improve the recognition and tracking performance because of the built-in interactions, as well as the temporal relationships among levels.

Once the image estimates are obtained, we can use them as the evidence to infer the true states of hidden nodes by maximizing the posterior probability as Eq. 3. Let \( E^t \), \( AU^t_{1:N} \), \( X^t_{\text{Feature}} \) (\( \text{Feature} \) stands for Eyebrow, Eye, Nose, Mouth) represent the nodes for Expression, \( N \) target AUs and facial feature points at time \( t \). Given the available evidence until time \( t \): \( M^{t|t}_{AU_{1:N}}, M^{t|t}_{X_{\text{Feature}}} \), where \( M^{t|t}_{AU_{1:N}} \) indicates the estimates of \( N \) target AUs from time 1 to \( t \) while \( M^{t|t}_{X_{\text{Feature}}} \) represents the estimates of facial feature nodes from time 1 to \( t \), the probability \( P(E^t, AU^t_{1:N}, X^t_{\text{Feature}}|M^{t|t}_{AU_{1:N}}, M^{t|t}_{X_{\text{Feature}}}) \) can be factorized and computed via the probabilistic model by performing the DBN updating process as follows [42]:

1) Prediction: Given the estimated probability distribution \( P(E^{t-1}, AU^{t-1}_{1:N}, X^{t-1}_{\text{Feature}}|M^{t-1|t-1}_{AU_{1:N}}, M^{t-1|t-1}_{X_{\text{Feature}}}) \), which is already inferred at time step \( t - 1 \), we could calculate the predicted probability \( P(E^t, AU^t_{1:N}, X^t_{\text{Feature}}|M^{t|t-1}_{AU_{1:N}}, M^{t|t-1}_{X_{\text{Feature}}}) \) by using the standard BN inference algorithm, such as a version of junction tree algorithm [52].

2) Rollup: Remove time slice \( t - 1 \) and use the prediction \( P(E^t, AU^t_{1:N}, X^t_{\text{Feature}}|M^{t|t-1}_{AU_{1:N}}, M^{t|t-1}_{X_{\text{Feature}}}) \) for the \( t \) slice as the new prior.

3) Estimation: Add new observations at time \( t \) and calculate the probability distribution over the current state \( P(E^t, AU^t_{1:N}, X^t_{\text{Feature}}|M^{t|t}_{AU_{1:N}}, M^{t|t}_{X_{\text{Feature}}}) \). Finally, add the slice for \( t + 1 \).

This way, we obtain the posterior probability of each hidden node, given the observed measurements. Because of the recursive nature of the inference process as well as the simple network topology, the inference can be implemented rather efficiently.

V. EXPERIMENTS

The proposed model is evaluated on two databases, i.e., the extended Cohn-Kanade (CK+) database [46], and the M&M Initiative (MMI) facial expression database [53]. CK+ database has by 22% larger number of sequences and 27% more number of subjects as compared the original Cohn-Kanade (CK) database [44]. One significant benefit of CK+ database compared to CK database is that the emotion labels on CK+ database are revised, while before the emotion labels were those that the actors have been told to express. CK and CK+ databases have been widely used for evaluating facial activity recognition system. Using CK+ database has several advantages: this database demonstrates diversity over the subjects and it involves multiple-AU expressions. The results on the CK+ database will be used to compare with other published methods. Besides, in order to evaluate the generalization ability of the proposed model, we train the model on CK+ database and test on the M&M Initiative (MMI) facial expression database collected by Pantic et al. [53]. The MMI facial expression database is recorded in true color with a frame rate of 24 fps. The advantage of using this database is that it contains a large number of videos that display facial expressions with a neutral-apex-neutral evolution.

A. Evaluation on Extended Cohn–Kanade Database

We collect 309 sequences from 90 subjects that contain the major six expressions from the CK+ database, 227 sequences of which are labeled frame by frame in this paper (all facial feature points, AUs and expressions). We adopt leave-one-subject-out cross validation, and for each iteration, while the semantic dependencies of the facial action model are trained with all labeled training images, the dynamic dependencies are learnt only using the sequences containing frame by frame labels. Given the AU and facial feature points measurements, the proposed model recognizes all three level facial activities simultaneously through a probabilistic inference. In the following, we are going to demonstrate the performance for each level individually.

1) Facial Feature Tracking: We tracked the facial feature point measurements through an active shape model (ASM) based approach [8], which first searches each point locally and then constrains the feature points based on the ASM model, so that the feature points can only deform in specific ways found in the training data. The ASM model is trained using 500 keyframes selected from the training data, which are 8-bit gray images with 640 × 480 image resolution. All the 26 facial feature point positions are manually labeled in each
training image. For ASM analysis, the principal orthogonal modes in the shape model stand for 95% of the shape variation. Since the face region is normalized and scaled based on the detected eye positions, the tracking model is invariant to scale change. The trained ASM model performs well when the expression changes slowly, but may fail when there is a large and sudden expression change. At the same time, our model can detect AUs accurately, especially when there is a large expression change. The trained ASM model performs well when the detected eye positions, the tracking model is invariant to scale change, since the face region is normalized and scaled based on the training image. For ASM analysis, the principal orthogonal modes in the shape model stand for 95% of the shape variation.

To evaluate the performance of the tracking method, the distance error metric is defined per frame as: \( \| \hat{p}_i - \hat{p}_i \| / D_I(j) \), where \( D_I(j) \) is the interocular distance measured at frame \( j \), \( \hat{p}_i \) is the tracked position of feature point \( i \), and \( \hat{p}_i \) is the labeled position. By modeling the interaction between facial feature points and AUs, our model reduces the average feature tracking error from 3.31 percent for the baseline method to 2.59 percent for the proposed model, a relative improvement of 21.75 percent. We also make a comparison with the active appearance model (AAM). Lucey et al. [46] provided AAM model tracking results on the CK+ database, and we selected 20 feature points from [46] that we also tracked for the same subjects in this paper. The comparison is listed in Table III. From Table III we can see that, AAM model outperforms the ASM based tracking method [8], mainly because both shape and texture are combined with PCA into an AAM model; however, the proposed model still achieves the best performance.

To further demonstrate the tracking effectiveness of the proposed model, we downsampled the frequency rate of some sequences from the CK+ database so that the expression and facial feature points positions can change excessively in two consecutive frames. In this way, it is more challenging for the traditional tracking model to track the facial feature points. The average tracking error of 26 facial feature points for a sequence is shown in Fig. 11. From Fig. 11 we can see that, the performances of the baseline method and the proposed model are similar for most frames, except the frames after frame 11. We show the 10th and the 11th frames in the figure, and we can see that the baseline tracking method fails because it is based on local search, and it cannot track the sudden lips part movement in the 11th frame because of downsampling. At the same time, detected AU measurements with high confidence, e.g., AU12+AU25, provide a prior distribution for the mouth shape, e.g., the parameter of the model \( P(X_{Mouth}|AU12 = 1, AU25 = 1) \) follows a multi-Gaussian distribution. Hence, the proposed model outperforms the baseline method for facial feature tracking when there is a sudden expression change. To clearly illustrate the top-down information flow from AUs to facial feature points, we initialize all AU measurement nodes with ground truth, and then infer the feature points. Through this way, we further reduce the average tracking error to 2.46 percent. Therefore, we can conclude that the top-down information flow from AUs to facial feature points can indeed help refine the tracking measurements.

2) Facial Action Unit Recognition: Fig. 12 shows the AU recognition performance for generalization to novel subjects.
on the CK+ database by using AdaBoost classifier alone and using the proposed model, respectively. From Fig. 12 we can see that, the proposed system outperforms the AdaBoost classifier consistently. The average F1 measure (a weighted mean of the precision and recall) for all target AUs increases from 69.94 percent for AdaBoost to 76.36 percent for the proposed model. We made one tailed t-test (right-tail test) on the average F1 measure from the proposed model and the AdaBoost, and the p-value is 3.003 × 10^{-11}, which means the predicted results are statistically better than the measurements. The improvement mainly comes from the AUs that are hard to detect but have strong relationships with other AUs. To clearly demonstrate this point, we list the parameters of AU15 node and AU23 node (ignoring the dynamic dependency) in Table IV. From Table IV, we can see that, the co-occurrence of AU15 and AU17 is high when AU7 is absent, i.e., \( P(\text{AU15} = 1 | \text{AU7} = 1, \text{AU17} = 1) = 0.0989 \), and the co-occurrence of AU23 and AU24 is high when AU15 does not occur, i.e., \( P(\text{AU23} = 1 | \text{AU15} = 0, \text{AU24} = 1) = 0.9309 \). By encoding such relationships into the DBN, the F1 measure of AU15 is increased from 61.54 percent to 72.07 percent; the F1 measure of AU17 is increased from 72.12 percent to 81.08 percent; the F1 measure of AU23 increases from 40.93 percent to 54.98 percent, and that of AU24 increases from 47.96 percent to 61.03 percent. Besides the semantic relationships among AUs, the interactions between AUs and facial feature points also contribute to the AU recognition. For instance, we initialize all facial feature measurements with ground truth, and then infer the AU nodes. In this way, the average F1 measure of AUs is further improved to 77.03 percent.

Since the AU labels for the overlap between CK and CK+ are exactly the same, as well as most previous works about AU recognition are evaluated on CK database, we make a comparison with some earlier works as listed in Table V. Our results in terms of classification rate are better than most previous works. Bartlett et al. [14] and Lucey et al. [38] both achieve high AU recognition rates, but these two approaches are all image-based, which usually evaluate only on the initial and peak frames while our method is sequence based and we consider the whole sequence, in the middle of which AUs with low intensity are much more difficult to recognize. In addition, the classification rate is often less informative, especially when the data is unbalanced. So we also report our results in terms of F1 measure, which is a more comprehensive metric. From Table V we can see that, the proposed method outperforms all the three earlier works who also reported their results in F1 measure. Since the works in [47] and [45] recognize more AUs, we also make a deep comparison on each individual AU as shown in Table VI. On average, our method achieves better or similar results, but it is interesting that for AU15 and AU24, our results are much better than the work in [47] and [19]. This is because the activations of AU15 and AU24 involve changes in facial texture without large displacements of feature points, and Valstar & Pantic employed geometric features in [47] and [19]. Hence, they failed at AU15 and AU24. The proposed approach also outperforms [19], [47] at AU9, the occurrence of which also produces less displacement change. P. Lucey et al. [46] provided the AU recognition results on the peak frames on the CK+ database, and for the same 15 AUs as recognized in this paper, [46] achieves an average area underneath the ROC curve of 89.41% for the similarity normalized shape features (SPTS), 91.27% for the canonical normalized appearance (CAPP) features and 93.92% for SPTS+CAPP features. The proposed model achieves an average area underneath the ROC curve of 93.33% for the peak frames, which is better or similar as that in [46].

**TABLE IV**

<table>
<thead>
<tr>
<th>Parameters of AU15</th>
<th>Parameters of AU23</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(\text{AU15} = 1</td>
<td>\text{AU7} = 1, \text{AU17} = 1) = 0.0989 )</td>
</tr>
<tr>
<td>( P(\text{AU15} = 1</td>
<td>\text{AU7} = 1, \text{AU17} = 0) = 0.0022 )</td>
</tr>
<tr>
<td>( P(\text{AU15} = 1</td>
<td>\text{AU7} = 0, \text{AU17} = 1) = 0.7096 )</td>
</tr>
<tr>
<td>( P(\text{AU15} = 1</td>
<td>\text{AU7} = 0, \text{AU17} = 0) = 0.0025 )</td>
</tr>
</tbody>
</table>

**TABLE V**

<table>
<thead>
<tr>
<th>Author</th>
<th>Features</th>
<th>Classification</th>
<th>AUs</th>
<th>CR</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chang 2006 [36]</td>
<td>manifold embed</td>
<td>Bayesian</td>
<td>23</td>
<td>89.4</td>
<td></td>
</tr>
<tr>
<td>Whitehill and Omlin 2006 [37]</td>
<td>Haar wavelets</td>
<td>AdaBoost</td>
<td>11</td>
<td>92.4</td>
<td></td>
</tr>
<tr>
<td>Lucey et al. 2007 [38]</td>
<td>AAM</td>
<td>SVM</td>
<td>15</td>
<td>95.5</td>
<td></td>
</tr>
<tr>
<td>Pantic et al. 2006 [21]</td>
<td>tracked feature points</td>
<td>temporal rule-based</td>
<td>21</td>
<td>93.3</td>
<td></td>
</tr>
<tr>
<td>Valstar et al. 2006 [19]</td>
<td>tracked feature points</td>
<td>AdaBoost+SVM</td>
<td>15</td>
<td>90.2</td>
<td>72.9</td>
</tr>
<tr>
<td>Koelstra et al. 2010 [45]</td>
<td>FFD</td>
<td>GentleBoost+HMM</td>
<td>18</td>
<td>89.8</td>
<td>72.1</td>
</tr>
<tr>
<td>Valstar &amp; Pantic 2012 [47]</td>
<td>tracked feature points</td>
<td>GentleSVM+HMM</td>
<td>22</td>
<td>91.7</td>
<td>59.6</td>
</tr>
<tr>
<td>This paper</td>
<td>Gabor filter, feature points</td>
<td>AdaBoost+DBN</td>
<td>15</td>
<td>94.05</td>
<td>76.36</td>
</tr>
</tbody>
</table>

AUs = No. of AUs recognized, CR = Classification Rate, F1 = F1 measure
TABLE VI
COMPARISON WITH SOME PREVIOUS WORKS ON INDIVIDUAL AUs ON CK/CK+ DATABASE

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.93</td>
<td>82.6</td>
<td>86.89</td>
</tr>
<tr>
<td>2</td>
<td>80.11</td>
<td>83.3</td>
<td>90.00</td>
</tr>
<tr>
<td>4</td>
<td>77.48</td>
<td>63.0</td>
<td>73.13</td>
</tr>
<tr>
<td>5</td>
<td>63.55</td>
<td>59.6</td>
<td>80.00</td>
</tr>
<tr>
<td>6</td>
<td>77.11</td>
<td>80.0</td>
<td>80.00</td>
</tr>
<tr>
<td>7</td>
<td>62.41</td>
<td>29.0</td>
<td>46.75</td>
</tr>
<tr>
<td>9</td>
<td>78.84</td>
<td>57.3</td>
<td>77.27</td>
</tr>
<tr>
<td>12</td>
<td>89.99</td>
<td>83.6</td>
<td>83.72</td>
</tr>
<tr>
<td>15</td>
<td>70.27</td>
<td>36.1</td>
<td>70.27</td>
</tr>
<tr>
<td>17</td>
<td>81.08</td>
<td></td>
<td>76.29</td>
</tr>
<tr>
<td>24</td>
<td>60.13</td>
<td>44.0</td>
<td>63.16</td>
</tr>
<tr>
<td>25</td>
<td>88.19</td>
<td>74.8</td>
<td>95.60</td>
</tr>
<tr>
<td>27</td>
<td>95.52</td>
<td>85.4</td>
<td>87.50</td>
</tr>
<tr>
<td>Avg</td>
<td>77.26</td>
<td>61.59</td>
<td>77.74</td>
</tr>
</tbody>
</table>

F1 = F1 measure of our model
F1 [47] = F1 Valstar & Pantic 2012 [47]
F1 [45] = F1 Koelstra et al. 2010 [45]

TABLE VII
EXPRESSION RECOGNITION CONFUSION MATRIX OF THE PROPOSED MODEL ON CK+ DATABASE

<table>
<thead>
<tr>
<th></th>
<th>Surp</th>
<th>Hap</th>
<th>Dis</th>
<th>Fear</th>
<th>Sad</th>
<th>Ang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surp</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Hap</td>
<td>0%</td>
<td>97.08%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Dis</td>
<td>0%</td>
<td>0%</td>
<td>91.02%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Fear</td>
<td>20.00%</td>
<td>0%</td>
<td>0%</td>
<td>80.00%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Sad</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>80.00%</td>
<td>0%</td>
</tr>
<tr>
<td>Ang</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>66.67%</td>
</tr>
</tbody>
</table>

Average Recognition Rate: 87.43%

Surp = Surprise, Hap = Happiness, Dis = Disgust, Sad = Sadness, Ang = Anger

3) Expression Recognition: Besides more accurate facial feature tracking and AU recognition, our model recognizes six global expressions with an average recognition rate of 87.43%. The result is not as good as that of the state-of-the-art expression recognition methods, e.g., [31], [39]. This is mainly because we have not employed any image-based methods specifically to obtain expression estimates, and instead the expression states are directly inferred from facial feature point and AU estimates, and from their relationships. Table VII shows the confusion matrix for six expressions recognition on the CK+ data set. From Table VII we can see that, the recognition rate for surprise and happiness are high while that of anger is low. This is mainly because we infer expressions from the corresponding AUs, and AU1, AU2, AU27 for surprise and AU6, AU12, AU25 for happiness are well detected. Hence, we can recognize these two expressions with high accuracy. At the same time, AUs for anger, i.e., AU5, AU7, AU23 and AU24, are all not detected with such high accuracy, so we only achieve a recognition rate of 66.67% for anger. Hence, we can conclude that the accuracy of the AU detection affects the expression recognition significantly in this model. To further demonstrate this point, we initialize all AU nodes with ground truth, and then infer the expression. We achieve an average expression recognition rate of 95.15% in this case, which is similar as that of the state-of-the-art methods in [31] (95.1%) and [39] (94.48%).

Besides, our approach allows a probabilistic output for six expressions, which represents the confidence of the classification and can be further transferred into the relative intensity level. Fig. 13 shows the expression recognition results of a sequence from CK+ database, in which the subject is performing surprise expression.

B. Generalization Validation Across Different Databases

In order to evaluate the generalization ability of the proposed model, we train the model on the extended Cohn-Kanade database and test on the MMI facial expression database [53]. Since most of the image sequences on the MMI database have only single AU active, we only choose 54 sequences containing two or more target AUs from 11 different subjects. The proposed model achieves an average expression recognition rate of 82.4%, and reduces the average tracking error from 3.96 percent for the baseline method [8] to 3.51 percent for the proposed model, an relative improvement of 11.36%. Fig. 14 shows the AU recognition results of using AdaBoost classifier alone and using the DBN facial action model, respectively, on the MMI database. From Fig. 14 we can see that, AU9 and AU15 on the MMI database are not well recognized. This is mainly because on the MMI database, these two actions occur rarely (less than 5%), and the appearance changes caused by these two actions are relatively dissimilar as that on the CK+ database. In addition, the co-occurrence of AU15 and AU17 on the MMI database is not as strong as that on the CK+ database, which is crucial for our model to improve the AU recognition performance on AU15. On average, with the use of the facial action model,
we improve the F1 measure of AU recognition from 61.97 percent for the AdaBoost, to 66.52 percent for the proposed model. The most current works by Valstar and Pantic. [47] and Koelstra et al. [45], which represent the state of the art methods for AU recognition, reported an average F1 measure of 53.79 percent and 65.70 percent respectively on the MMI database. The model of the same 15 AUs as recognized in this paper.

We improve the F1 measure of AU recognition from 61.97 percent for the AdaBoost, to 66.52 percent for the proposed model. The most current works by Valstar and Pantic. [47] and Koelstra et al. [45], which represent the state of the art methods for AU recognition, reported an average F1 measure of 53.79 percent and 65.70 percent respectively on the MMI database. The proposed model achieves better AU recognition performance than the state of the art methods [45], [47] on novel subjects from a different database, which demonstrates the generalization ability of our model.

The enhancement of our approach mainly comes from combining the facial action model with image-based methods. Specially, the erroneous image measurement could be compensated by the semantic and dynamic relationships encoded in the DBN. For instance, the recognition of AU7 is difficult since the contraction of AU7 produces a similar facial appearance changes as that caused by AU6. However, AU7 occurs often with AU4, which could be recognized easily. By encoding such co-occurrence relationship in the DBN model, the F1 measure of AU7 is increased greatly (from 61.22 percent to 70.82 percent). Similarly, by modeling the co-occurrence relationships of AU23 and AU24, the F1 measure of AU23 is increased from 58.72 percent to 76.34 percent, and that of AU24 is increased from 75.25 percent to 83.02 percent.

VI. CONCLUSION

In this paper, we proposed a hierarchical framework based on Dynamic Bayesian Network for simultaneous facial feature tracking and facial expression recognition. By systematically representing and modeling inter relationships among different levels of facial activities, as well as the temporal evolution information, the proposed model achieved significant improvement for both facial feature tracking and AU recognition, compared to state of the art methods. For six basic expressions recognition, our result is not as good as that of state of the art methods, since we did not use any measurement specifically for expression, and the global expression is directly inferred from AU and facial feature point measurements and from their relationships. The improvements for facial feature points and AUs come mainly from combining the facial action model with the image measurements. Specifically, the erroneous facial feature measurements and the AU measurements can be compensated by the model’s build-in relationships among different levels of facial activities, and the build-in temporal relationships. Since our model systematically captures and combines the prior knowledge with the image measurements, with improved image-based computer vision technology, our system may achieve better results with little changes to the model.

In this paper, we evaluate our model on posed expression databases from frontal view images. In the future work, we plan to introduce the rigid head movements, i.e., head pose, into the model to handle multi view faces. In addition, modeling the temporal phases of each AU, which is important for understanding the spontaneous expression, is another interesting direction to pursue.

ACKNOWLEDGMENT

This project was funded in part by a scholarship from China Scholarship Council (CSC). The work was accomplished when the first author visited Rensselaer Polytechnic Institute (RPI) as a visiting student. We would like to acknowledge support from CSC and RPI.

REFERENCES

K. Schmidt and J. Cohn, “Dynamics of facial expression: Normative

S. W. Chew, R. Rana, P. Lucey, S. Lucey, and S. Sridharan, “Sparse Tem-

Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. I. Sejnowski,


Y. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial

G. Zhao and M. Pietikainen, “Boosted multi-resolution spatiotemporal
descriptors for facial expression recognition,” Pattern Recognit.

Valstar and M. Pantic, “Combined support vector machines and
hidden Markov models for modeling facial action temporal dynamics,”
pp. 118–127.

Valstar and M. Pantic, “Fully automatic facial action unit detection and

K. B. Korb and A. E. Nicholson, Bayesian Artificial Intelligence.

M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and
J. Movellan, “Automatic recognition of facial actions in spontaneous

T. Kanade, J. Cohn, and Y. L. Tian, “Comprehensive database for facial
expression analysis,” in Proc. 4th IEEE Int. Conf. Autom. Face Gesture

S. Koelstra, M. Pantic, and I. Patras, “A dynamic texture-based approach

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews,
“The extended Cohn-Kande dataset (CK+): A complete facial expression
dataset for action unit and emotion-specified expression,” in Proc. 3rd

M. Valstar and M. Pantic, “Fully automatic recognition of the temporal

I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for
tracking facial features,” in Proc. IEEE Int. Conf. Autom. Face Gesture

H. Dibeklioglu, A. A. Salah, and T. Gevers, “A statistical method for
2-D facial landmarking,” IEEE Trans. Image Process., vol. 21, no. 2,

I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, “Facial expres-
sion recognition from video sequences: Temporal and static modeling,”

Y. Zhang and Q. Ji, “Active and dynamic information fusion for facial
expression understanding from image sequences,” IEEE Trans. Pattern

U. Kjærulff, “dHugin: A computational system for dynamic time-sliced

for facial expression analysis,” in Proc. IEEE Int. Conf. Multimedia


Yongqiang Li received the B.S. and M.S. degrees in
instrument science and technology from the Harbin
Institute of Technology, Harbin, China, in 2007 and
2009, respectively, where he is currently pursuing the
Ph.D. degree.

He was a Visiting Student with Rensselaer Poly-
technic Institute, Troy, NY, USA, from 2010 to 2012.
His current research interests include computer
vision, pattern recognition, and human-computer
interaction.

Shangfei Wang (M’02) received the M.S. degree in
electrical engineering and the Ph.D. degree in signal
and information processing from the University of
Science and Technology of China (USTC), Hefei,
China, in 1999 and 2002, respectively.

She was a Post-Doctoral Research Fellow with
Kyushu University, Japan, from 2004 to 2005. She is
currently an Associate Professor with the School of
Computer Science and Technology, USTC. Her
current research interests include computation intel-
ligence, affective computing, multimedia computing,
information retrieval and artificial environment design.
Yongping Zhao received the Ph.D. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China. He is currently a Professor with the Department of Instrument Science and Technology, Harbin Institute of Technology. His current research interests include signal processing, system integration, and pattern recognition.

Qiang Ji received his Ph.D. 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 recently served as a program director at the National Science Foundation (NSF), where he managed NSF’s computer vision and machine learning programs. He 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 Dept. of Computer Science at 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.

Prof. Ji’s research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published over 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including NSF, NIH, DARPA, ONR, ARO, and AFOSR as well as by major companies including Honda and Boeing. He is an editor on several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. He is a fellow of IAPR and a senior member of the IEEE.