Simultaneous Facial Feature Tracking and Facial Expression Recognition

Yongqiang Li, Yongping Zhao, Shangfei Wang, and Qiang Ji

Abstract

The tracking and recognition of facial activities from images or videos 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 (AUs), defined in Facial Action Coding System, represent the contraction of a specific set of facial muscles, i.e., lid tightener, 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 emotion state. 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 (DBN) 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

Simultaneous tracking and recognition, facial feature tracking, facial action unit recognition, expression recognition, Bayesian network.

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 the facial activities in three levels. First, in the bottom level, facial feature tracking,

Y. Li, S. Wang, and Q. Ji are with the Rensselaer Polytechnic Institute, Troy, NY, USA; Y. Zhao is with the Harbin Institute of Technology, Harbin, China.
Email: {liy23, wangs9, jiq}@rpi.edu.

which usually detects and tracks prominent 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 FACS [1], try to recognize some meaningful facial activities (i.e., lid tightener, eyebrow raiser, etc); In the top level, facial expression analysis attempts to recognize facial expressions that represent the human emotion 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, the facial feature tracking can be used in the feature extraction stage in expression/AUs recognition, and the expression/AUs recognition results can provide a prior distribution for the 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 computer vision measurements 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 approach, 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/AUs recognition help 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 (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 reviews on the related works on facial activity analysis; Sec. III describes the details of facial activity modeling,

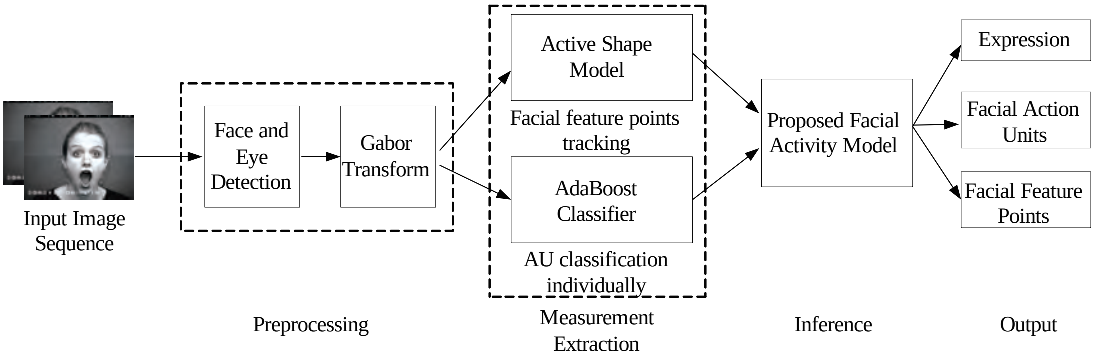


Fig. 1. The flowchart of the online facial activity recognition system

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 faical action model; Sec. V shows the experimental results on two databases. The paper concludes in Sec .VI with a summary of our work and its future extensions.

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 is 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 [49] [50] [51] are general purpose point trackers without the prior knowledge of the object. Each facial 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 explicit 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 that are 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 [2] are subsequently proposed to combine constraints of both shape variation and texture variation.

In the conventional statistical models, i.e. ASM, the feature point positions are updated (or projected) simultaneously, which indicates that the interactions within feature points are simply concurrent. 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 (AUs) congenitally 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 in computer vision on facial expression analysis [22] [14] [9] [16] [25]. Current methods in this area can be grouped into two categories: image-driven method and model-based method.

Image-driven 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 (i.e., length of wrinkles and degree of eye opening) [16], Haar features [11] [38], Local Binary Patterns (LBP) features [32] [33], independent component analysis (ICA) [12], feature points [49], 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] [21], rule-based approach [22], AdaBoost classifiers, Sparse Representation (SR) classifiers [34] [35], etc. A survey about expression recognition can be found in [23].

The common weakness of appearance-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, i.e., [46] [17]. Model-based methods overcome this weakness by making use of the relationships among AUs, and recognize the AUs simultaneously. Lien et al. [24] 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 Hidden Markov Model (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. They, however, fail to exploit the spatial dependencies among AUs. To remedy this problem, Tong and Ji [26] [25] employed a Dynamic Bayesian network to systematically model the spatiotemporal relationships among AUs, and achieved significant improvement over the image-driven method. In this work, 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 [52] [49] [53] [48], and integrating face tracking with video coding [28]. However, in most of these works, the interaction between facial feature tracking and facial expression recognition is one-way, i.e., feed facial feature tracking results to facial expression recognition [49] [53]. There is no feedback from the recognition results to facial feature tracking. Most recently, Dornaika et al. [27] and Chen & Ji [31] improved the facial feature tracking performance by involving the facial expression recognition results. However, in [27], they only model six expressions and they need to retrain the model for a new subject, while in [31], 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:

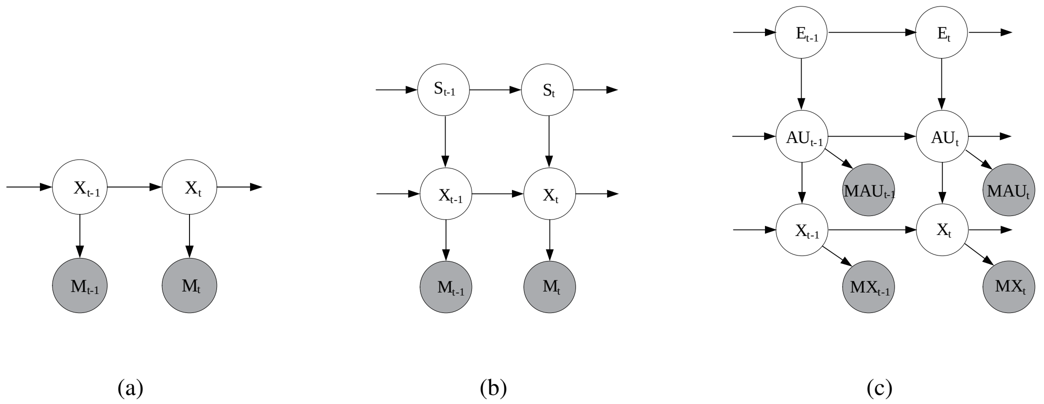


Fig. 2. Comparison of different tracking models: (a) traditional tracking model, (b) tracking model with switch node, (c) and the proposed facial activity tracking model.

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 AU 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, i.e., facial feature points, we want to track, and M_t is the current image measurement (Hereafter, the shaded nodes represent measurements and the unshaded nodes denote the hidden states). The directed links are quantified by the conditional probabilities, i.e. 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.

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

$M_{1:t}$ is the measurement sequence from frame 1 to t. If both X_t and M_t are continuous and all the condition 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 underling dynamic system [29] [30]. For the switching dynamic model, the switch node represents different states and for each state, there are particular predominant movement patterns. The works in [27] and [31] 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:

$$\begin{aligned} P(X_t, S_t|M_{1:t}) \propto P(M_t|X_t) \int_{X_{t-1},S_{t-1}} & P(X_t|X_{t-1}, S_t) \\ & P(S_t|S_{t-1})P(X_{t-1}, S_{t-1}|M_{1:t-1}) \end{aligned} \tag{2}$$

In [27], they propose 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; AU_t represents a set of AUs; X_t denotes the facial feature points we are going to track; MAU_t and MX_t 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 facial feature point 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 psychologist

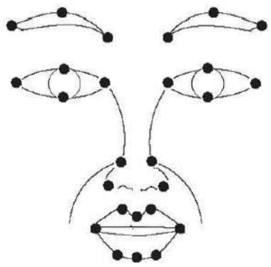


Fig. 3. Facial feature points we tracked.

agree presently that there are ten basic emotions, most current research in facial expression recognition mainly focuses on six major emotions, partially because they are the most basic, and culture 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:

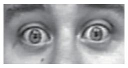
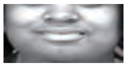
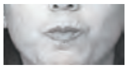
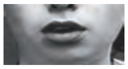
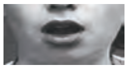
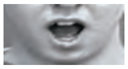
$$E_t^\star, AU_t^\star, X_t^\star = argmax_{E_t, AU_t, X_t} \\ P(E_t, AU_t, X_t | MAU_{1:t}, MX_{1:t}) \tag{3}$$

B. Modeling the Relationships between Facial Features and AUs

In this work, 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 27 as summarized in Table I. The selection of AUs to recognize is mainly based on the AUs occurrence frequency, their importance to characterize the 6 expression, and the amount 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 adequate training data set. Facial action units control the movement of face component 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 facial feature points around each facial component to represent the casual relationships between them. Take $Mouth$ for example, we use a continuous node X_{Mouth} to represent 8 facial points around mouth, and link AUs that control mouth movement to this

TABLE I

A LIST OF AUS AND THEIR INTERPRETATIONS

AU1 Inner brow raiser	AU2 Outer brow raiser	AU4 Brow Lowerer	AU5  Upper lid raiser	AU6 Cheek raiser
AU7 Lid tigherner	AU9 Nose wrinkler	AU12  Lip corner puller	AU15 Lip corner depressor	AU17 Chin raiser
AU23  Lip tigherner	AU24 Lip presser	AU25  Lip part	AU26  Jaw Drop	AU27  Mouth stretch

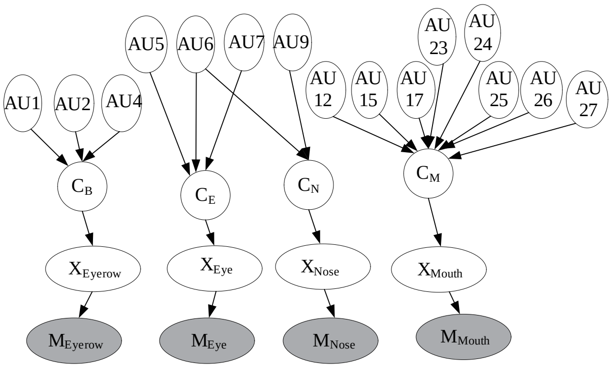


Fig. 4. (a) Modeling the relationships between facial feature points and AUs ($C_{B/E/N/M}$ are the intermediate nodes; $X_{Eyebrow/Eye/Nose/Mouth}$ are the facial points nodes around each face component and $M_{Eyebrow/Eye/Nose/Mouth}$ are the corresponding measurement nodes).

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 AU or AU combinations for the mouth. Thus, only a set of common AU or AU combination, which produce significant facial actions, is sufficient to control the face component movement. As a result, we introduce an intermediate node, i.e., “C''_M to model the correlations among AUs and to reduce the number of AU combination. 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 modeling of the semantic relationships among AUs will be discussed in the later section. 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 component. 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 $P(C_B = 1|AU1 = 1, AU2 = 1, AU4 = 1) = 0.9$ to represent that the eyebrow is entirely raised up.

The facial feature points nodes (i.e., $X_{Eyebrow}$, X_{Eye}, X_{Nose} and X_{Mouth}) have continuous state and are represented by continuous shape vectors. 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:

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

with the mean shape vector μ_k and covariance matrix Σ_k.

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

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

with the mean shape vector μ_x, regression matrix W, and covariance matrix Σ_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, i.e., AU1+AU2+AU5+AU26+AU27 to show surprise expression; AU6+AU12+AU25 to show happiness expression; AU1+AU4+AU15+AU17 to show sadness expression, etc.

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 work, we propose to learn from the data to capture additional relationships among AUs.

Tong et al. [26] 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-driven methods. Following the work in [26], 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 work, and the dependency is characterized by a conditional probability table (CPT), i.e., θ, for each 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 work, we employ the Bayesian Information Criterion (BIC) score function [41] which is defined as follows:

$$s_D(G) = \max_{\theta} \log P(D|G,\theta) - \frac{\log M}{2} Dim_G \tag{6}$$

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 θ with respect to data D and structure G; M is the number of training data; and Dim_G is the number of parameters.

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

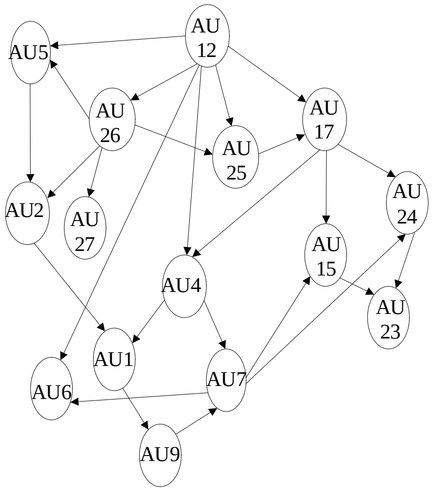


Fig. 5. The learned structure from training data.

TABLE II

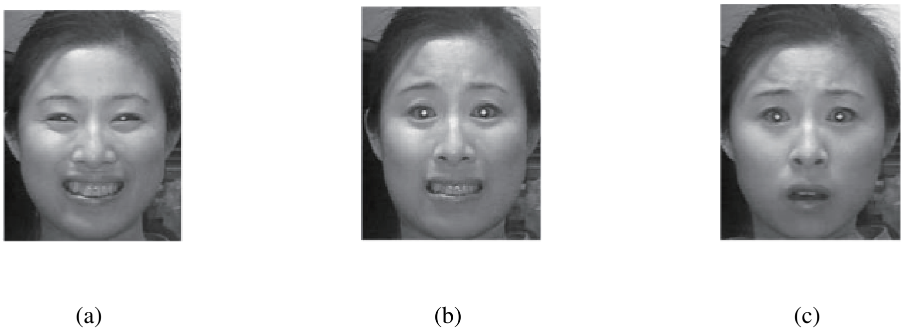
GROUPING AUS ACCORDING TO DIFFERENT EXPRESSIONS.

Emotion	Corresponding AUs
Surprise	AU5, AU26, AU27, AU1+AU2
Happiness	AU6, AU12, AU25
Sadness	AU1, AU4, AU15, AU17
Disgust	AU9, AU17
Anger	AU4, AU5, AU7, AU23, AU24
Fear	AU4, AU1+AU5, AU5+AU7

D. Modeling the Relationships between AUs and Expression

So far, we have modeled the relationships between AUs and facial feature points, and the semantic relationships among AUs. 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, i.e., 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, i.e., when AU1 occurs alone, it indicates a sadness,



(a)

(b)

(c)

Fig. 6. Examples of AU combinations: (a) AU12+AU6 (two AUs from the same category) enhances classification to happiness, (b) AU1+AU5 (two AUs from different categories) becomes a fear, and (c) AU26+AU1 (two AUs from different categories) increases ambiguity between a surprise and a fear.

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, i.e., 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 denpendency, 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

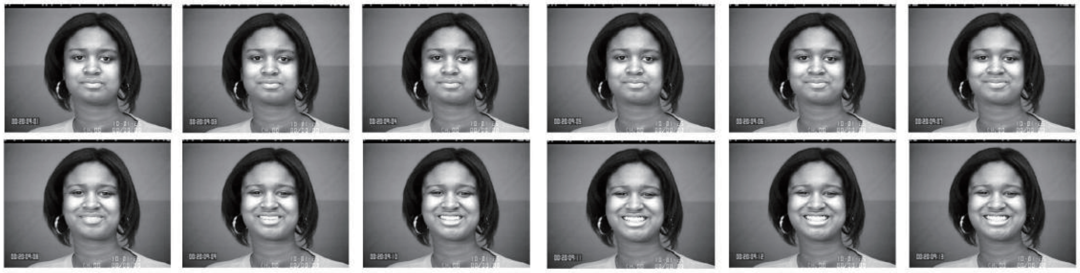


Fig. 7. An example image sequence displays the unsynchronized AUs evolutions in a smile (adapted from [45]).

resolution of 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, i.e., a blink, to several minutes, i.e., a jaw clench, as described in [22]. Hence, if we choose a small time duration, i.e., 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 and the analysis of the database, as well as the temporal characteristics of the AUs we intend to recognize, we empirically set the time duration as 1/6 second in this work, and link $AU2$ and $AU12$ at time $t-1$ to $AU5$ and $AU6$ at time t, respectively to capture the second type dynamics.

In the proposed framework, we consider two types of conditional dependencies for variables at two adjacent time slices. The first type, i.e., an arc from AU_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 point nodes, we only consider this type dynamic. The second of type, i.e., an arc from AU_i at time $t-1$ to $AU_j (j \neq i)$ at time t, depicts how AU_i at the previous time step affects $AU_j (j \neq i)$ 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. [36] 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

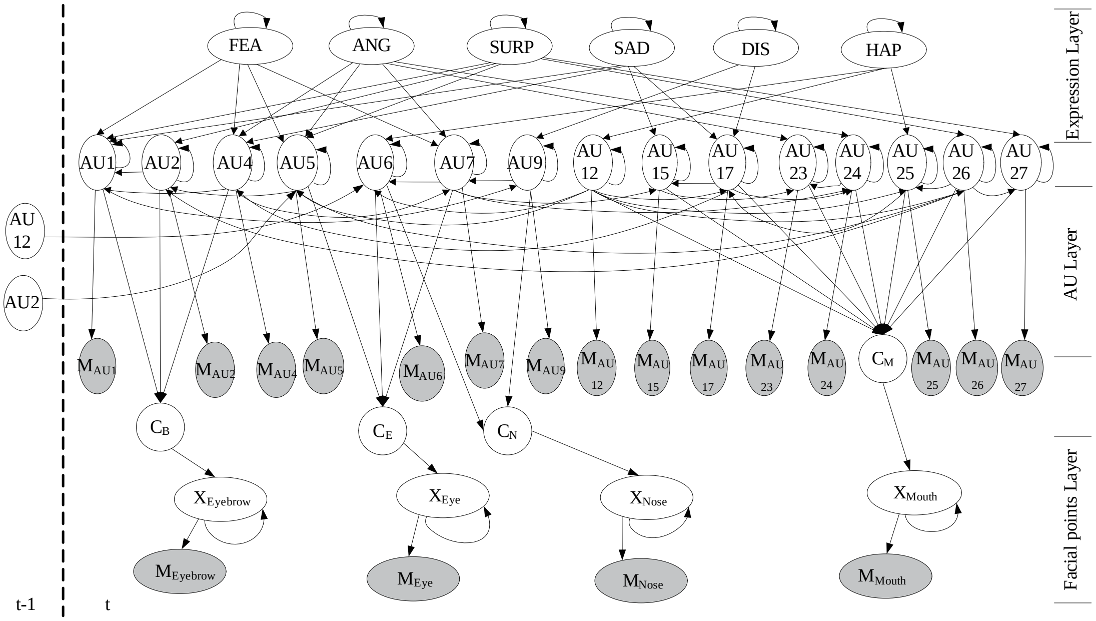


Fig. 8. The complete DBN model for simultaneous facial activity recognition. The shaded node indicates the observation for the connected hidden node. The self-arrow at the hidden node represents its temporal evolution from previous time slice to the current time slice. The link from AU_i at time $t-1$ to $AU_j (j \neq i)$ at time t indicates the dynamic dependency between different AUs.

first action unit to follow AU12 in 47%. Similar findings are found by Tong et al [20]. 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 [47] frame by frame. 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 for a larger network structure. For instance, for the AU and expression model, since all nodes are discrete and let θ_{ijk} indicates a probability parameter,

$$\theta_{ijk} = p(x_i^k | pa^j(X_i)) \tag{7}$$

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 kth state of variable X_i, and $pa^j(X_i)$ is the jth configuration of the parent nodes of X_i. The "fitness" of parameters θ and training data D is quantified by the log likelihood function $log(p(D|\theta))$, denoted as $L_D(\theta)$:

$$L_D(\theta) = log \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}} \tag{8}$$

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. Since we have complete training data, the learning process can be described as a constrained optimization problem as follows:

$$arg \max_{\theta} L_D(\theta) \qquad s.b. \qquad g_{ij}(\theta) = \sum_{k=1}^{r_i} \theta_{ijk} - 1 = 0 \tag{9}$$

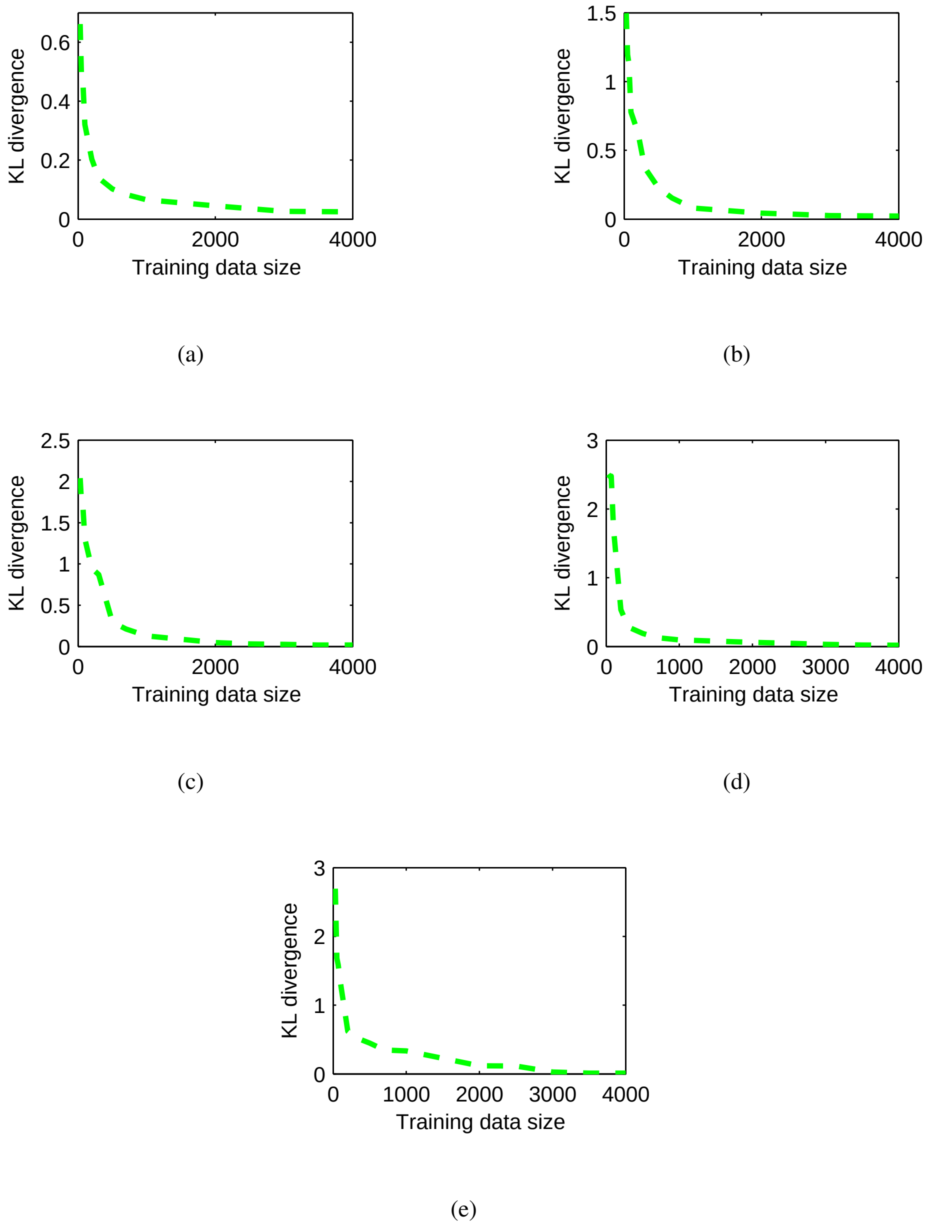
Solving the above equations, we can get $\theta_{ijk} = \frac{n_{ijk}}{\sum_k n_{ijk}}$.

For the facial feature point model, i.e., the $Mouth$ model, we need to learn a mean shape vector and a covariance matrix for each state of the combination node. Since the combination node is hidden, in this work, 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 amount 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: $P(X_{Mouth}|C_M)$ as shown in Fig. 10 (The X_{Mouth} node in our model is the shape difference. For clarity, we show the distribution of X_{Mouth} by adding a constant neutral shape: $P(X_{Mouth} + C|C_M)$, where C is a constant neutral shape). From Fig. 10 we can observe that, AUs do 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



(a)

(b)

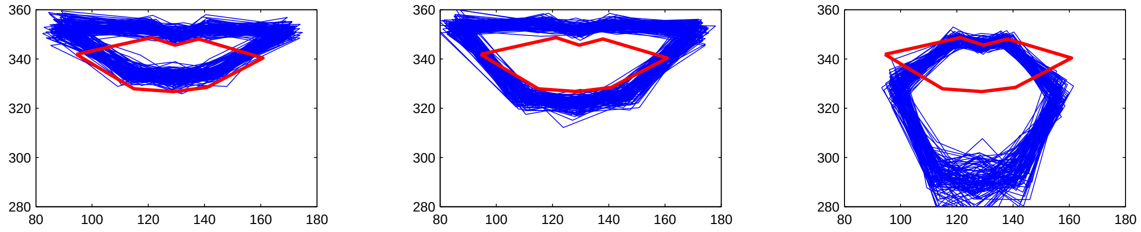
(c)

(d)

(e)

Fig. 9. The KL divergences of the model parameters versus the training data size for the (a) AU model, (b) "Eyebrow" model, (c) "Eye" model, (d) "Nose" model, (e) and "Mouth" model respectively.

likelihood of the measurement to estimate the posterior probability. Therefore, the estimation contains two steps in our framework. First, we employ various computer vision techniques to acquire various measurements. For AUs, we employ a technique based on the AdaBoost classifier and Gabor features [44] to obtain AU measurements. For facial feature points, we first use the detection method [8] to obtain the facial 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



(a) (b) (c)

Fig. 10. The distribution of mouth feature points given different AUs. (a) $P(X_{Mouth}|AU12 = 1)$ (b) $P(X_{Mouth}|AU12 = 1, AU25 = 1)$ (c) $P(X_{Mouht}|AU25 = 1, AU27 = 1)$. The red shape indicates the mouth neutral shape.

wavelet matching and active shape model. In this work, we infer expressions directly from the corresponding AUs, which means we do not employ any image-driven method to obtain the expression measurements first. Without measurement, 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 measurements are obtained, we can use them as the evidence to estimate the true state of hidden nodes by maximizing the posterior probability of the hidden nodes as Eq. 3. Let E^t, $AU^t_{1:N}$, $X^t_{Feature}$ ($Feature$ represents $Eyebrow, Eye, Nose, Mouth$), represent the nodes for Expression, N target AUs and facial feature points at time t. Given the available evidence until time t: $MAU^{1:t}_{1:N}$, $M^{1:t}_{X_{Feature}}$, the probability $p(E^t, AU^t_{1:N}, X^t_{Feature}|MAU^{1:t}_{1:N}, M^{1:t}_{X_{Feature}})$ can be factorized and computed via the facial action model by performing the DBN updating process as follows [43]:

1) Prediction: Given the estimated probability distribution $p(E^{t-1}, AU^{t-1}_{1:N}, X^{t-1}_{Feature}|MAU^{1:t-1}_{1:N}, M^{1:t-1}_{X_{Feature}})$, which is already inferred at time step $t-1$, we could calculate the predicted probability $p(E^t, AU^t_{1:N}, X^t_{Feature}|MAU^{1:t-1}_{1:N}, M^{1:t-1}_{X_{Feature}})$ by using the standard BN inference algorithm, such as a version of junction tree algorithm [54].
2) Rollup: Remove time slice $t-1$ and use the prediction $p(E^t, AU^t_{1:N}, X^t_{Feature}|MAU^{1:t-1}_{1:N}, M^{1:t-1}_{X_{Feature}})$ 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 $p(E^t, AU^t_{1:N}, X^t_{Feature}|MAU^{1:t}_{1:N}, M^{1:t}_{X_{Feature}})$. 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, it can be implemented rather efficiently.

V. Experiments

The proposed model is evaluated on two databases, i.e., the extended Cohn-Kanade (CK+) database [47], and the M&M Initiative (MMI) facial expression database [55]. CK+ database increases the original Cohn-Kanade (CK) database [45] by 22% in the number of sequence, and by 27% in the number of subjects. 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. [55]. 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 that contain the major six expressions from the CK+ database, 227 sequences of which are labeled frame by frame in this work. 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 point 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 facial feature point locally and then constrains the feature point positions based on the ASM model, so that the facial feature points can only deform in specific ways found in the training data. The ASM

TABLE III

ERRORS OF TRACKING FACIAL FEATURE POINTS BY USING BASELINE METHOD [8], AAM MODEL [47] AND THE PROPOSED MODEL, RESPECTIVELY. (FOR AAM MODEL, WE SELECTED 20 POINTS FROM [47] THAT WE ALSO TRACKED IN THIS WORK.)

	Eyebrow	Eye	Nose	Mouth	Total
Baseline method [8]	3.75	2.43	3.10	3.97	3.31
AAM model [47]	3.43	2.36	2.76	3.65	3.05
Proposed model	2.98	1.53	2.43	3.45	2.59

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 and not significantly, 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 accurately detected AUs provide a prior distribution for the facial feature points, which help infer the true point position.

To evaluate the performance of the tracking method, the distance error metric is defined per frame as: $\frac{||p_{i,j}-\widehat{p}_{i,j}||_2}{D_I(j)}$, where $D_I(j)$ is the interocular distance measured at frame j, $p_{i,j}$ is the tracked position of point i, and $\widehat{p}_{i,j}$ is the labeled position. By modeling the interaction between facial feature points and AUs, our model reduces the average facial feature points 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., [47] provided AAM model tracking results on the CK+ database, and we selected 20 points from [47] that we also tracked for the same subjects in this work. 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 that both shape and texture are combined with PCA to one AAM model, and 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 point positions can change significantly in two consecutive frames. In this way,

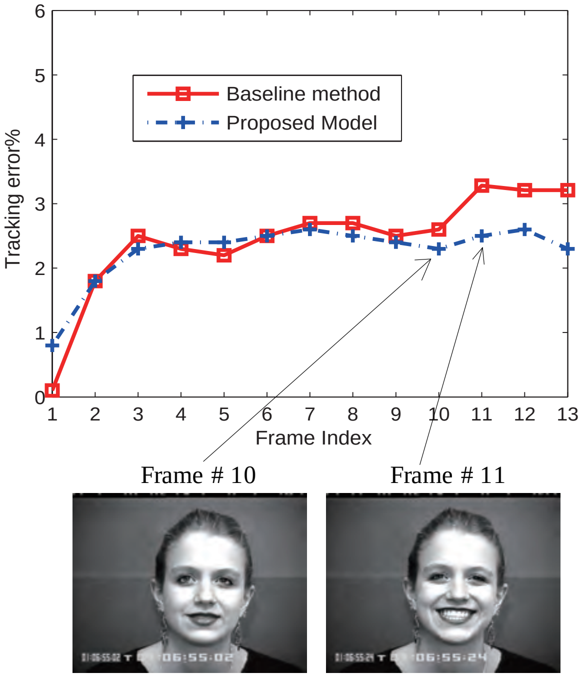


Fig. 11. The tracking error for 26 facial feature points of baseline method [8] and the proposed model.

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, i.e., AU12+AU25, provide a prior distribution for the mouth shape, i.e., the parameter of the model $P(X_{Mouth}|AU12 = 1, AU25 = 1)$ follows a multi-gaussian distribution as shown in Fig. 10(b). 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 facial 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

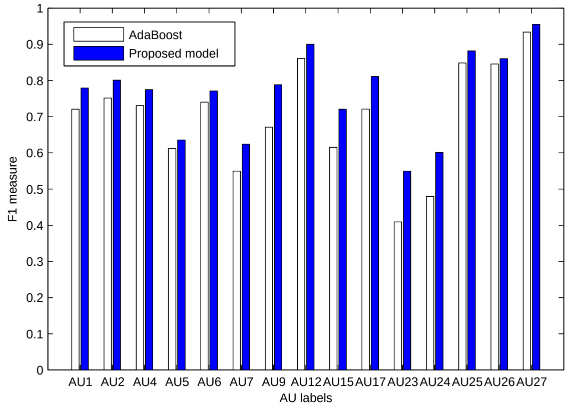


Fig. 12. Comparison of AU recognition results on the novel subjects on CK+ database by using AdaBoost classifier and using the proposed model, respectively.

TABLE IV

Model parameters of AU15 node and AU23 node (ignoring the dynamic dependency).

Parameters of AU15	Parameters of AU23
$P(AU15=1\|AU7=1,AU17=1)=0.0989$	$P(AU23=1\|AU15=1,AU24=1)=0.0883$
$P(AU15=1\|AU7=1,AU17=0)=0.0002$	$P(AU23=1\|AU15=1,AU24=0)=0.0416$
$P(AU15=1\|AU7=0,AU17=1)=0.7096$	$P(AU23=1\|AU15=0,AU24=1)=0.9309$
$P(AU15=1\|AU7=1,AU17=1)=0.0025$	$P(AU23=1\|AU15=1,AU24=1)=0.0052$

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) respectively in Table IV. From Table IV, we can see that, when AU7 is absent the co-occurrence of AU15 and AU17 is high, i.e., $P(AU15=1|AU7=0,AU17=1)=0.7096$, and when AU15 does not occur the co-occurrence of AU23 and AU24 is high, i.e., $P(AU23=1|AU15=0,AU24=1)=0.9309)$. By encoding such relationships in 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

TABLE V

COMPARISON OF OUR WORK WITH SOME PREVIOUS WORKS.

Author	features	classification	AUs	CR	F1
Bartlett et al. 2005 [14]	Gabor filters	AdaBoost+SVM	17	94.8	
Chang 2006 [37]	manifold embed	Bayesian	23	89.4	
Whitehill and Omlin 2006 [38]	Haar wavelets	AdaBoost	11	92.4	
Lucey et al. 2007 [39]	AAM	SVM	15	95.5	
Pantic et al. 2006 [22]	tracked face points	temporal rule-based	21	93.3	
Valstar et al. 2006 [20]	tracked face points	AdaBoost+SVM	15	90.2	72.9
Tong el al. 2007 [26]	Gabor filters	AdaBoost+DBN	14	93.3	
Koelstra el al. 2010 [46]	FFD	GentleBoost+HMM	18	89.8	72.1
Valstar & Pantic 2012 [49]	tracked facial points	GentleSVM+HMM	22	91.7	59.6
This work	Gabor filter, face points	AdaBoost+DBN	15	94.05	76.36

AUs = No. of AUs recognized, CR = Classification Rate, F1 = F1 measure

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.

Lots of works about AUs recognition are evaluated on CK, or CK+ databases. Table V shows the comparison of the proposed model with some earlier works. Our results in terms of classification rate are better than most previous works. Bartlett et al. [14] and Lucey et al. [39] all achieve high AU recognition rate, 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 [49] and [46] 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 [49] and [20]. This is because the activations of AU15 and AU24 involve changes in facial texture without large displacements of facial feature points, and Valstar & Pantic employed geometric feature in [49] and [20]. Hence, they failed at AU15 and AU24. The proposed approach also outperforms [49] [20] at AU9, the occurrence of which also produces

TABLE VI

COMPARISON WITH SOME PREVIOUS WORKS ON INDIVIDUAL AUS.

AUs	F1	F1[49]	F1[46]	F1[20]
1	77.93	82.6	86.89	87.6
2	80.11	83.3	90.00	94.0
4	77.48	63.0	73.13	87.4
5	63.55	59.6	80.00	78.3
6	77.11	80.0	80.00	88.0
7	62.41	29.0	46.75	76.9
9	78.84	57.3	77.27	76.4
12	89.99	83.6	83.72	92.1
15	70.27	36.1	70.27	30.0
17	81.08		76.29	
24	60.13	44.0	63.16	14.3
25	88.19	74.8	95.60	95.3
27	95.52	85.4	87.50	89.3
Avg	77.26	61.59	77.74	75.80

F1 = F1 measure of our model

F1 [49] = F1 Valstar & Pantic 2012[49]

F1 [46] = F1 Koelstra el al. 2010[46]

F1 [20] = F1 Valstar & Pantic 2006[20]

less displacement change. P. Lucey et al. [47] provided the AU recognition results on the peak frames on the CK+ database, and for the same 15 AUs as recognized in this work, [47] 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 [47].

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 the-state-of-the-art expression recognition methods, i.e., [32] [40]. This is mainly because that, we didn't employ any image-driven method specifically to extract the expression measurement, and its state is directly inferred from facial feature point and AU measurements, 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 that we infer expressions from the corresponding AUs, and AU1, AU2, AU27 for surprise

TABLE VII

EXPRESSION RECOGNITION CONFUSION MATRIX OF THE PROPOSED MODEL.

	Surp	Hap	Dis	Fear	Sad	Ang
Surp	**96.88%**	0%	0%	3.12%	0%	0%
Hap	0%	**97.08%**	0%	0%	2.92%	0%
Dis	0%	0%	**91.02%**	0%	8.98%	0%
Fear	20.00%	0%	0%	**80.00%**	0%	0%
Sad	0%	0%	0%	0%	**80.00%**	20.00%
Ang	0%	0%	8.33%	0%	25.00%	**66.67%**
					Average Recognition Rate:	**87.43%**

Surp = Surprise, Hap = Happiness, Dis = Disgust

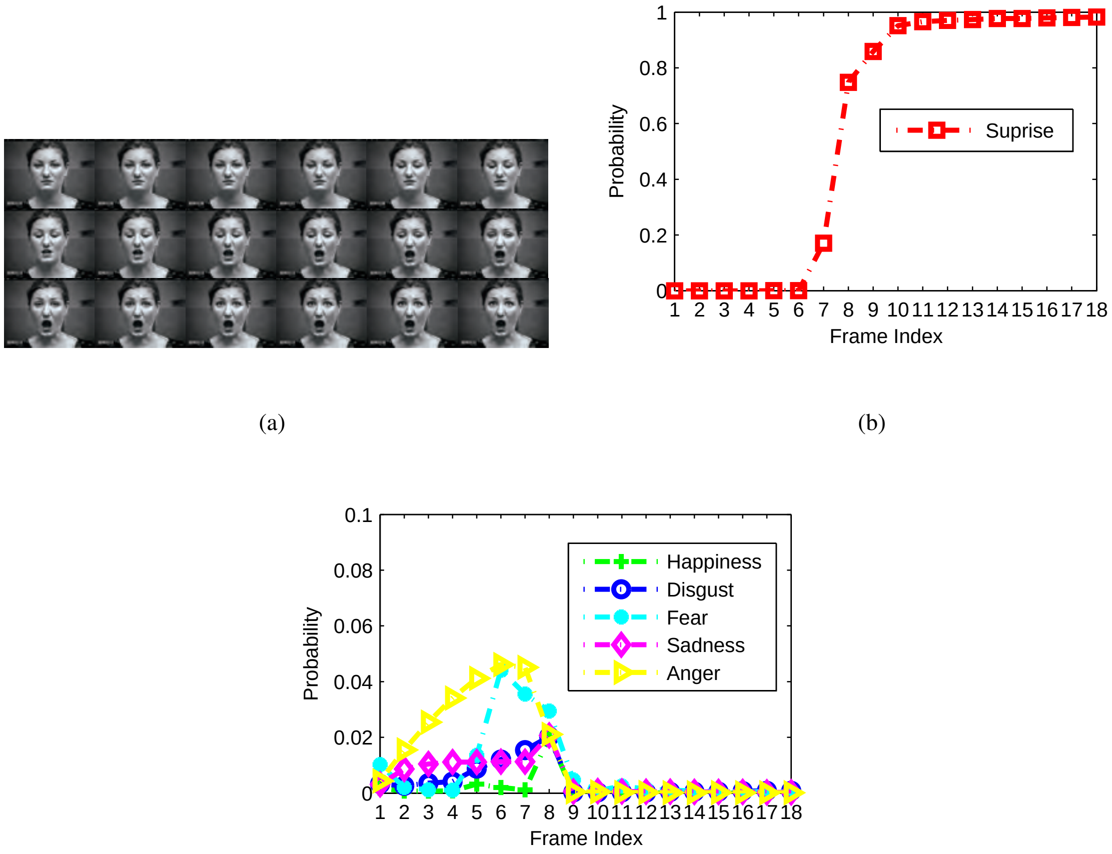
Sad = Sadness, Ang = Anger

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 the state of the art method in [32](95.1%) and [40](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. 12 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 [55]. 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 reduce 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



(a)

(b)

(c)

Fig. 13. Expression recognition results on a sequence. (a) A sequence on CK+ database and the subject is performing surprise expression. (b) The corresponding recognition results of surprise. (c) The corresponding recognition results for other five expressions.

MMI database. With the use of the facial action model, we improve the average 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 Vasltar and Pantic. [49] and Koelstra et al. [46], 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[1]. The proposed model achieves better AU recognition performance than the state of the art methods [49] [46] on novel subjects from a different database, which demonstrates the generalization ability of our model.

The enhancement of the our approach mainly comes from combing the facial action model with image driven methods. Specially, the erroneous image measurement could be compen-

[1] For work [49], we calculate the average F1 measure of the same 13 AUs as recognized in this work, while for work [46], we calculate the average F1 measure of the same 15 AUs as recognized in this work.

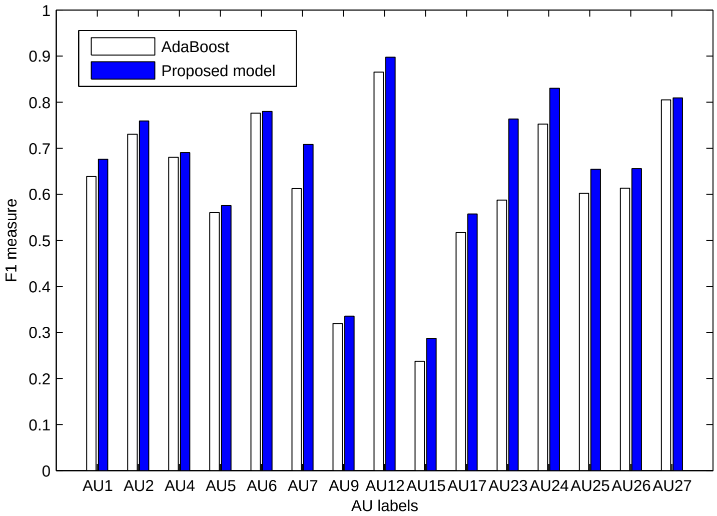


Fig. 14. AU recognition results on MMI facial expression database by using AdaBoost classifier and using the proposed model, respectively. The model is trained on CK+ database and tested on MMI database.

sated 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 other 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 relationship 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 driven 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.

REFERENCES

[1] P. Ekman and W. V. Friesen, Facial Action Coding System (FACS): Manual, Consulting Psychologists Press, 1978.

[2] Z. Zhu, Q. Ji, K. Fujimura, and K. Lee, Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination, in Proc. IEEE Intl Conf. Pattern Recognition, pp. 318-321, 2002.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, Active shape modelsłtheir training and application, Comput. Vision Image Understanding, vol. 61, no. 1, pp. 38-59, 1995.

[4] T. F. Cootes, G. J. Edwards, and C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell, vol. 23, no. 6, pp. 681-685, 2001.

[5] X. W. Hou, S. Z. Li, H. J. Zhang, and Q. S. Cheng, Direct appearance models, in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 828-833, 2001.

[6] S. J. McKenna, S. Gong, R. P. Wrtz, J . Tanner, and D. Banin, Tracking facial feature points with Gabor wavelets and shape models, in Proceedings of International Conference on Audio- and Video-based Biometric Person Authentication, pp. 35-42, 1997.

[7] M. Rogers and J. Graham, Robust active shape model search, in Proceedings of ECCV, pp. 517-530, 2002.

[8] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji, Robust facial feature tracking under varying face pose and facial expression, Pattern Recognition, 2007.

[9] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li, Detection, Tracking, and Classification of Action Units in Facial Expression, J. Robotics and Autonomous System, vol. 31, pp. 131-146, 2000.

[10] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, Classifying Facial Actions, IEEE Trans. Pattern Anal. Mach. Intell, vol. 21, no. 10, pp. 974-989, Oct. 1999.

[11] R. Lienhart and J. Maydt, An extended set of Haar-like features for rapid object detection, in Proc. IEEE Intl Conf. Image Processing, 2002.

[12] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, Recognizing faces with PCA and ICA, Computer Vision and Image Understanding, vol. 91, pp. 115-137, 2003.

[13] C. P. de Campos and Q. Ji, Efficient structure learning of bayesian networks using constraints, Journal of Machine Learning Research, pp. 663-689, 2011.

[14] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 568-573, 2005.

[15] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, Classifying Facial Actions, IEEE Trans. Pattern Anal. Mach. Intell, pp. 974-989, Oct. 1999.

[16] Y. Tian, T. Kanade and J. F. Cohn, Recognizing Action Units for Facial Expression Analysis, IEEE Trans. Pattern Anal. Mach. Intell, vol. 23, no. 2, pp. 97-115, 2001.

[17] G. Zhao and M. Pietikainen, Boosted Multi-Resolution Spatiotemporal Descriptors for Facial Expression Recognition, Pattern Recognition Letters, vol. 30, no. 12, pp. 1117-1127, 2009.

[18] M. Valstar and M. Pantic, Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics, Lecture Notes on Computer Science, pp. 118-127, 2007

[19] M. Valstar and M. Pantic, Fully Automatic Recognition of the Temporal Phases of Facial Actions, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, 2012

[20] M. Valstar and M. Pantic, Fully Automatic Facial Action Unit Detection and Temporal Analysis, in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2006.

[21] A. Kapoor, Y. Qi, and R. W. Picard, Fully Automatic Upper Facial Action Recognition, in Proc. IEEE Intl Workshop Analysis and Modeling of Faces and Gestures, pp. 195-202, 2003.

[22] M. Pantic and I. Patras, Dynamics of Facial Expressions-Recogniton of Facial Actions and Their Temporal Segments from Face Profile Image Sequences, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 2, pp. 433C449, 2006.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Trans. Pattern Anal. Mach. Intell, vol. 31, no. 1, 2009.

[24] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li, Detection, Tracking, and Classification of Action Units in Facial Expression, Robotics and Autonomous Systems, 1999.

[25] Y. Tong, J. Chen, and Q. Ji, A Unified Probabilistic Framework for Spontaneous Facial Activity Modeling and Understanding, IEEE Trans. Pattern Anal. Mach. Intell, vol. 32, no. 2, 2010.

[26] Y. Tong, W. Liao, and Q. Ji, Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships, IEEE Trans. Pattern Anal. Mach. Intell, vol. 29, no. 10, 2007.

[27] F. Dornaika and F. Davoine, Simultaneous facial action tracking and expression recognition in the presence of head motion, International Journal of Computer Vision, 2008.

[28] K. Schwerdt and J. L. Crowley, Robust face tracking using color, in Proc. IEEE Intl Conf. Autom. Face Gesture Recog., 2000.

[29] Y. B. Shalom, and X. Li, Estimation, Tracking: Principles, Techniques, and Software, Hardcover, Artech House Publishers, 1993.

[30] Z. Ghahramani, and G. E. Hinton, Variational Learning for Switching State-Space Models, Neural Computation, vol. 12, 2000.

[31] J. Chen and Q. Ji, A Hierarchical Framework for Simultaneous Facial Activity Tracking, in Proc. IEEE Intl Conf. Autom. Face Gesture Recog., 2011

[32] C. Shan, S. Gong, and P. W. McOwan, Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study, Image and Vision Computing, 2009

[33] G. Zhao and M. Pietikainen, Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions, IEEE Trans. Pattern Anal. Mach. Intell, vol. 29, no. 6, pp. 915-928, 2007.

[34] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, Facial Action Unit Recognition with Sparse Representation, in Proc. IEEE Intl Conf. Autom. Face Gesture Recog., 2011.

[35] S. W. Chew, R. Rana, P. Lucey, S. Lucey, and S. Sridharan, Sparse Temporal Representations for Facial Expression Recognition, Lecture Notes in Computer Science, 2012.

[36] K. Schmidt and J. Cohn, Dynamics of Facial Expression: Normative Characteristics and Individual Differences, in Proc. IEEE Intl Conf. Multimedia and Expo, pp. 728-731, 2001.

[37] Y. Chang, C. Hu, R. Feris, and M. Turk, Manifold-Based Analysis of Facial Expression, J. Image and Vision Computing, pp. 605-614, 2006

[38] J. Whitehill and C. W. Omlin, Haar Features for FACS AU Recognition, in Proc. IEEE Intl Conf. Autom. Face Gesture Recog., 2006

[39] S. Lucey, A. Ashraf, and J. Cohn, Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, Face Recognition, K. Delac and M. Grgic, eds., pp. 275-286, I-Tech Education and Publishing, 2007

[40] L. Zhang and D. Tjondronegoro, acial Expression Recognition Using Facial Movement Features, IEEE Trans. Affective Computing, 2011

[41] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics, vol. 6, pp. 461-464, 1978.

[42] D. Heckerman, D. Geiger, and D. M. Chickering, Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Machine Learning, vol. 20, no. 3, pp. 197-243, 1995.

[43] K. B. Korb and A. E. Nicholson, Bayesian Artificial Intelligence, Chapman and Hall/CRC, 2004.

[44] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, Automatic recognition of facial actions in spontaneous expressions, Journal of Multimedia, 2006.

[45] T. Kanade, J. Cohn, and Y. L. Tian, Comprehensive database for facial expression analysis, in Proc. IEEE Intl Conf. Autom. Face Gesture Recog., 2000.

[46] S. Koelstra, M. Pantic, and I. Patras, A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models, IEEE Trans. Pattern Anal. Mach. Intell, vol. 32, 2010.

[47] 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. IEEE Intl Conf. Comput. Vis. Pattern Recog., Workshop, 2010

[48] H. C. Akakin and B. Sankur, Robust classification of face and head gestures in video, Image and Video Computing, vol. 29, pp. 470-483, 2011.

[49] M. Valstar and M. Pantic, Fully Automatic Recognition of the Temporal Phases of Facial Actions, IEEE Trans. Syst., Man, Cybern. B, Cybern., pp. 28-43, 2012.

[50] I. Patras and M. Pantic, Particle filtering with factorized likelihoods for tracking facial features, in Proc. IEEE Int'l Conf. Autom. Face Gesture Recog., pp. 97-102, 2004, .

[51] H. Dibeklioglu and A. A. Salah, A Statistical Method for 2-D Facial Landmarking, IEEE trans. Image Processing, vol. 21, 2012

[52] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huanga, Facial expression recognition from video sequences: temporal and static modeling, Computer Vision and Image Understanding, vol. 91, 2003.

[53] Y. Zhang and Q. Ji, Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences, IEEE Trans. Pattern Anal. Mach. Intell, vol. 27, no. 5, pp. 699-714, 2005.

[54] U. Kjaerulff, dHugin: A Computational System for Dynamic Time-Sliced Bayesian Networks, Intl J. Forecasting: Special Issue on Probability Forecasting, vol. 11, pp. 89-111, 1995.

[55] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, Web-Based Database for Facial Expression Analysis, in Proc. IEEE Intl Conf. Multimedia and Expo, July 2005.