

A Hierarchical Framework for Simultaneous Facial Activity Tracking

Jixu Chen Qiang Ji

Department of Electrical, Computer and System Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180

chenj4@rpi.edu qji@ecse.rpi.edu

Abstract—The tracking of facial activities from video is an important and challenging problem. Nowadays, many computer vision techniques have been proposed to characterize the facial activities in the following three levels (from local to global): First, in the bottom level, the facial feature tracking focuses on detecting and tracking the prominent local landmarks surrounding facial components (e.g. mouth, eyebrow, etc); Second, the facial action units (AUs) characterize the specific behaviors of these local facial components (e.g. mouth open, eyebrow raiser, etc); Finally, facial expression, which is a representation of the subjects' emotion (e.g. Surprise, Happy, Anger, etc.), controls the global muscular movement of the whole face. Most of the existing methods focus on one or two levels of facial activities, and track (or recognize) them separately.

In this paper, we propose to exploit the relationships among multi-level facial activities and track the facial activities in the three levels simultaneously. Specifically, we propose a unified stochastic framework based on the Dynamic Bayesian network (DBN) to explicitly represent the facial involvements in different levels, their interactions and their observations. By modeling the relationships among the three level facial activities, the proposed method can improve the tracking (or recognition) performance in all three levels.

I. INTRODUCTION

Facial activity, which is the major source of information for understanding emotional state and intention, has drawn growing attention in industry and academia. In recent years, plenty of the computer vision techniques have been developed to track or recognize the facial activities in three levels. In the bottom level, we can capture a detailed face shape by tracking the facial feature points, which are the prominent landmarks surrounding facial components. But sometimes we are only interested in the higher level information, such as some meaningful facial behaviors, e.g. mouth open, eyebrow raiser. etc. Based on the psychological studies of Ekman's facial action coding system (FACS) [12], we can use the facial action units (AUs) resulting from the local facial muscular movements to characterize these facial behaviors. In the top level, facial expression analysis attempt to recognize six basic facial expressions, i.e., happy, surprise, sadness, fear, disgust and anger [12]. These basic expressions represent the global facial muscular movement.

Accurate localization and tracking local facial feature points is important in the applications such as animation and human-machine interaction. In general, the facial feature tracking techniques can be classified into model-free and model-based algorithms. The model-free algorithms only use the general purpose point trackers [24], [5]. However, the

point tracker are susceptible to the inevitable errors due to noise or occlusion. Recently, extensive work has been focused on the model-based facial feature tracking which utilized the facial shape constraints, such as active shape model (ASM) [9], active appearance model (AAM) [8] and elastic bunch graph matching (EBGM) [29].

In the literature, the facial expression recognition systems are usually focused on either the recognition of the six typical expressions [6], [11], [31], [7], [22] or the recognition of the AUs [16], [10], [4], [1], [2], [26], [27]. Since the temporal facial involvement brings more information of the expression/AUs, most attention has been given to the temporal approach which try to recognize the expression/AUs in the video sequence. In general, an expression recognition system consists of two key stages: First, various facial features are extracted to represent the facial gestures or facial movements, e.g. dense optical flow are used in [16], [10] to detect the direction and magnitude of the facial movements; Bartleet et al. [1] convolves the whole face image by a set of Gabor wavelet kernels, and the resulting Gabor wavelet magnitude are used as facial features. Given the extracted facial features, the expression/AUs are identified by recognition engines, such as the Neutral Networks [10], [4], [23], Hidden Markov Models [16], Adaboost classifier [1], [2] and Bayesian networks [7], [31], [26], [27].

The facial feature tracking, AU recognition and expression recognition represent the facial activity in three levels, and they are interdependent problems. For example, the facial feature tracking can be used in the feature extraction stage in expression recognition, and the expression recognition result can provide a prior shape information in the model-based facial feature tracking. However, most current systems only track the facial activities in one or two levels, and track them separately, ignoring their interactions. In addition, the computer vision measurements in each level are always uncertain and ambiguous. They are uncertain because of the presence of noise, occlusion, and of the imperfect nature of the vision algorithm. They are ambiguous because they only measure certain aspects of the visual activity, e.g. the facial feature tracking usually depends on local search and it is prone to drift, on the other hand, the expression recognition depends on global features but lose some details. Therefore, one expects to better infer the facial activities by systematically combining the measurements from multiple sources.

The idea of combining tracking with other problem has been attempted before, such as simultaneous face tracking and recognition [33], and integrating face tracking with video coding[21]. Most recently, Fadi et al [11] proposed a simultaneous facial action tracking and expression recognition algorithm. In their algorithm, they track the deformation of a 3D face mesh. By utilizing the dynamic of the expression and modeling the relationships between the expression and the 3D face mesh, the tracking performance is improved. However, since their model is subject-dependent, they need to train the model for each subject. Furthermore, they only model six basic expressions which is a very small subset of human expressions.

We propose a unified probabilistic framework based on the dynamic Bayesian network to explicitly model the relationships among the three level facial activities and their dynamics. This framework can also be seen as an information fusion process that combine the measurements from multiple levels. Finally, the expression, AU and facial features are recovered simultaneously through a probabilistic inference.

Zhang et al. use a DBN for expression recognition [31]. But they perform facial feature tracking and expression recognition sequentially, i.e. the extracted facial feature is the input to expression recognition system. In our simultaneous tracking and recognition model, we explicitly model the interactions between facial feature tracking and expression recognition, i.e. the detected expression (or AUs) can also improve the facial feature tracking result. Furthermore, instead of setting the model parameters manually, we learn the DBN parameters automatically from large face database, which includes various expressions and subjects.

II. OVERVIEW OF THE FACIAL TRACKING MODELS

From the tracking point of view, our method is an extension to the traditional tracking model (e.g., Kalman Filter), which only model one facial dynamics, and the simultaneous tracking model in [11] which models multiple facial dynamics.

A. Tracking with Single Dynamics

The graphical model representation of the traditional tracking algorithm is shown in Fig.1(a). X_t is the current hidden state we want to track, I_t is the current image measurement. (Hereafter, the shaded nodes denote the measurement nodes and the unshaded nodes represent the hidden nodes). The directed links represent the conditional probabilities, i.e. the link from X_t to I_t represents the likelihood $P(I_t|X_t)$ and the link from X_{t-1} to X_t represents the first order dynamics $P(X_t|X_{t-1})$.

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

$$P(X_t|I_{1:t}) \propto P(I_t|X_t) \int_{X_{t-1}} P(X_t|X_{t-1})P(X_{t-1}|I_{1:t-1}) \quad (1)$$

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

B. Simultaneous Tracking with Multiple Dynamics

The above tracking model only have one single dynamics ($P(X_t|X_{t-1})$), and this dynamics is fixed for the whole sequence. However, face usually exhibits complex and rich dynamic behaviors. So the better modeling of dynamics can provide a powerful cue in the presence of measurement noise. Most recently, Fadi et. al [11] proposed to model the facial dynamics for different expressions. Their idea can be interpreted as the graphical model in 1(b). (Note that, the model in [11] actually uses second order dynamics, i.e. there is a link from X_{t-2} to X_t . To represent the basic idea of that paper, we only show this simplified first-order dynamic model). X_t is the ‘‘facial action’’ which denotes the movements of facial feature points, E_t represents the six basic expressions, and I_t is the image measurement.

This facial tracking model is similar to the Switching Linear Dynamics System (SLDS) [19], where the high level state controls the underlying dynamic system. Specifically, the dynamics of the facial feature ‘‘switches’’ according to the expression, i.e. the dynamics from X_{t-1} to X_t depends on the current expression E_t . Correspondingly, we can see that X_t has both X_{t-1} and E_t as its parents in 1(b) and its dynamic is $P(X_t|X_{t-1}, E_t)$.

Through this model, X_t and E_t can be tracked simultaneously, and their posterior probability is:

$$P(X_t, E_t|I_{1:t}) \propto P(I_t|X_t) \cdot \int_{X_{t-1}, E_{t-1}} P(X_t|X_{t-1}, E_t) \cdot P(E_t|E_{t-1}) \cdot P(X_{t-1}, E_{t-1}|I_{1:t-1}) \quad (2)$$

In [11], they propose to use particle filtering to estimate this posterior probability.

The above two models track only one or two levels of the facial activities. And note that, although the algorithm in [11] track the facial feature and the expression simultaneously, they only have the measurement for facial features. The expression doesn’t have direct measurement and its state is estimated from the lower facial feature level.

III. DBN MODEL FOR MULTI-LEVEL FACIAL TRACKING

Dynamic Bayesian Network (DBN) [18] is a directed graphical model, which models the temporal evolution of a set of random variables. Compared with the tracking models in section II, DBN is more general and can capture complex relationships among the variables.

We propose to use the DBN model in Figure 1(c) to track the three levels of facial activities simultaneously. The E_t node in the top level represents the current expression; UA_t and LA_t represent the AUs related to the upper-face and the AUs related to the lower-face, respectively; UX_t and LX_t node denotes the facial feature points on the upper-face and the facial feature points on the lower-face respectively. UAM_t , LAM_t , UXM_t and LXM_t represent the corresponding measurements of AUs and facial feature points. We will describe the computer vision techniques to obtain these measurements in section IV. Note that, although there is no strong relationship (directly link) between the upper-face AU and the lower-face AU, it doesn’t mean that they are independent. Actually, based on the ‘‘d-separation’’ property

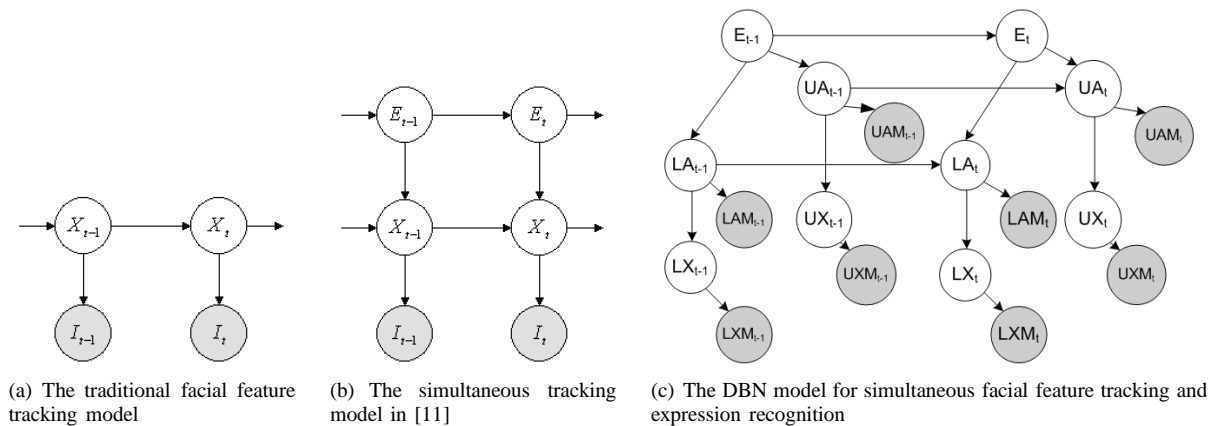


Fig. 1. Compare DBN with other facial tracking models.

[20], they are still dependent with each other through the hidden node E_t .

Given the measurement sequences, the posterior of the three level facial activities are estimated through the inference in DBN (section V). And the optimal states are tracked by maximizing this posterior:

$$E_t^*, UA_t^*, LA_t^*, UX_t^*, LX_t^* = \underset{E_t, UA_t, LA_t, UX_t, LX_t}{\operatorname{argmax}} P(E_t, UA_t, LA_t, UX_t, LX_t | UAM_{1:t}, LAM_{1:t}, UXM_{1:t}, LXM_{1:t}) \quad (3)$$

Compared with the traditional tracking model with single dynamics 1(a), our model can capture the relationships among different levels and recover the facial feature points and expressions simultaneously. Compared with simultaneous tracking model in 1(b), our model considers both the global and local expressions, and fuse the information (measurements) from multiple levels to improve the robustness of tracking and recognition.

Furthermore, since our model structure in each time slice forms a simple tree, the model learning and inference can be performed efficiently and accurately. For model learning, since the model structure is manually set, we only need to learn the model parameters, i.e. conditional probability distributions (CPDs) of the nodes. We will shown in Section III-B that the CPD of each node can be learned separately and the closed-form solution exists to learn these parameters. Thus, the model learning is straightforward and stable. For inference, we will shown in Section V that tracking can be solved efficiently as an filtering problem in DBN.

A. DBN Model Parameterization

Given the DBN model in Fig.1(c), we need to define the states for each node and the conditional probability distribution (CPD) associated with each node. The CPD defines conditional probability of each node given its parents $P(X|pa(X))$. Hereafter, $pa(X)$ is defined as the set of parent nodes of node X . In this section, we will define the CPD of each node, and the method to learn the parameters of each CPD is discussed in section III-B

1) *Expression Level*: In the top expression level, we want to model the six basic expressions. E_t is a discrete node which has 8 possible states: happiness, sadness, disgust,

surprise, anger, fear, neutral and “others”. The “others” state denotes all the expressions that cannot be explained by the basic expressions.

The CPD of the expression $P(E_t|E_{t-1})$ can be represented as a 8×8 transition matrix T whose entries $T_{e,e'}$ denotes the probability of the transition from expression e to e' . Although this matrix can be learned from training data, we have found that a general near-diagonal matrix works well for all the sequences. We set the diagonal elements to be close to one, and the rest percentage are equally distributed for other expressions. This matrix actually gives a higher probability to the current expression if it is same as the previous one.

2) *AU Level*: Our modeling of expression level is similar to the simultaneous tracking model in section II-B. However, the six typical expressions only describe the global facial activities, and they are only a small set of the complex face expressions. For example, the “surprise” implies a widely opened mouth and a raise of the eyebrows. But in applications, the subject may open mouth widely without raising eyebrows. In this case, the model in section II-B will encounter problems, because this expression is not included in the six basic expressions, and there is no facial feature dynamics defined for this expression.

It is generally believed that the expressions can be described linguistically using culture and ethnically independent AUs. Such AUs were developed by Ekman and Friesen in their FACS [12], where each AU is coded based on the local facial muscle involvements. For example, AU27 (mouth stretch) describes a widely open mouth, and AU4 (brow lowerer) makes the eyebrows lower and pushed together. In [31], they exploit some primary AUs which are highly related to the six basic facial expressions. Here we model 11 AUs from these primary AUs. They are listed in Figure 2. (We only use a subset of AUs because of the limitation of the training data. In Section. III-B. we use the Cohn-Kanade database [13] to train our model. These 11 AUs and their combination covers 91% of the data. Each of other AUs only appears in a few frames, which make it impossible to learn a probabilistic model.)

Based on our DBN model in figure 1(c), we first group

the AUs into “upper-face AUs (UA)”, including AU1, AU2, AU4, AU6 and “lower-face AUs (LA)”, including AU12, AU15, AU17, AU23, AU24, AU25 and AU27. The two AU groups capture the local facial behaviors of upper-face and low-face respectively.












					
Inner Brow Raiser	Outer Brow Raiser	Brower Lowerer	Cheek Raiser	Lip Corner Puller	Lip Corner Depressor
					
Chin Raiser	Lip Tightener	Lip Pressor	Lips part	Mouth Stretch	

Fig. 2. The list of AUs.

Each single AU has two discrete values: 0 and 1 which represents “presence” and “absence” respectively. However, if we directly stack these binary AUs into a vector and model the state of the node UA (or LA) with this vector, there will be too many possible states. For example, for the lower face AU, there will be 2^7 possible states, but most of them rarely occur in daily life, i.e. have too few examples in the training data to learn their probabilities. In this work, we select a few most frequent AU combinations as typical AUs, and use the typical AUs as the states of AU node. Specifically, LA has 8 states including AU25, AU25+AU12, AU25+AU27, AU12, AU17/AU15, AU23/AU24, “neutral” and “others” and UA has 6 states including AU6, AU6+AU4, AU1+AU2, AU1+AU4, “neutral” and “others”. (“+” means two AUs happen together, “/” means either of the two AUs happens.) Here, the “others” state denotes all the AU combinations that cannot be explained as typical AUs.

As shown in figure 1(c), the AU node (LA or UA) has two parents: the previous AU and the current expression. For instance, the CPD of the lower-face AU is presented as $P(LA_t|LA_{t-1}, E_t)$, i.e., the first order dynamics of the lower-face AU is dependent on the current expression E_t . More specifically, there is a 8×8 transition matrix of LA and a 6×6 transition matrix of UA for each expression.

Finally, the measurement nodes (LAM and UAM) for the AUs represent their observations obtained through some computer vision techniques (See section IV). The AU measurement has the same discrete states as its corresponding AU. So, the conditional probability $P(LAM|LA)$ is modeled as a conditional probability table (CPT) which is a 8×8 matrix. Similarly, the CPT of $P(UAM|UA)$ is a 6×6 matrix. And these conditional probabilities represent the measurement uncertainty with the computer vision techniques.

3) *Facial Feature Level*: In this work, we focus on the 14 facial features around the mouth and eyebrows (as shown in Fig.3), which have significant movement under different expressions. Because the eyebrow and mouth shape in neutral face is different for different subject, to eliminate the neutral shape variance we subtract the neutral shape from current shape, and model the shape difference. LX is a 16 dimensional vector which denotes the x,y differences of the

8 mouth points. Similarly, UX denotes differences of the 6 eye brow points.

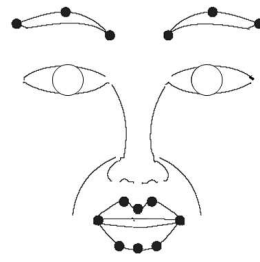


Fig. 3. The facial feature points round mouth and eyebrows.

Given the local AU, the CPD of facial feature points can be represented as a Gaussian distribution, e.g., for lower-face:

$$P(LX_t|LA_t = k) = \mathbf{N}(LX_t; \mu_k, \Sigma_k) \quad (4)$$

with the mean shape vector μ_k and covariance matrix Σ_k . Based on the conditional independence embedded in the BN, we could learn μ_k and Σ_k locally as shown in section III-B

Finally, the measurement nodes of the feature points represent their observations from computer vision techniques (See section IV). The facial feature measurements are the continuous vectors which have the same dimension as their parents. And CPD of measurement is modeled as a linear Gaussian distribution [17], e.g., for lower-face:

$$P(LXM_t|LX_t = lx) = \mathbf{N}(LXM_t; W_L \cdot lx + \mu_L, \Sigma_L) \quad (5)$$

with the mean shape vector μ_L , regression matrix W_L , and covariance matrix Σ_L . These parameters can be learned from the training data.

B. DBN Model Learning

Given the definition of the CPDs, we need to learn the parameters of the CPDs from training data. In this learning process, we manually labeled the expression, the AU, and the facial features in the training sequences. These labels are the ground-truth states of the hidden nodes. The states of measurement nodes are also obtained by applying various computer vision techniques in the training sequences (See section IV). Then we have the complete data for all the nodes. Based on the conditional independence embedded in the BN, we could learn each CPD locally.

The Cohn and Kanade’s DFAT-504 database [13] is used as training data. All the image sequences are coded into AUs frame by frame to learn the dynamics of AUs. We also manually label the facial feature points for each frame.

In the AU level, Table I shows the learned transition matrices of LA in the expression “happy” and “surprise”. For clarity, only a few representative elements of the whole 8×8 matrix are shown. First, because of the temporal smoothness, we can see that the self-transitions (diagonal elements) have higher probabilities. The other transitions can represent the specific dynamics for each expression. For instance, in “happy” expression, neutral face has high probability to transfer to AU12 (lip corner puller), and AU12 has high probability to transfer to AU25+AU12; in “surprise”

	Neutral	25	25+12	12
Neutral	0.5235	0.0157	0.0890	0.3717
25	0.0002	0.3633	0.6356	0.0002
25+12	0.0000	0.0000	0.9975	0.0024
12	0.0000	0.0000	0.1773	0.8227

(a)

	Neutral	25	25+12	25+27	12
Neutral	0.6022	0.2841	0.0000	0.1023	0.0057
25	0.0000	0.4347	0.0000	0.5651	0.0000
25+12	0.0010	0.0010	0.4971	0.4971	0.0010
25+27	0.0000	0.0000	0.0000	0.9999	0.0000
12	0.0006	0.0006	0.3323	0.0006	0.6639

(b)

TABLE I

(A) THE TRANSITION MATRIX OF LA IN “HAPPY” EXPRESSION. (EACH ENTRY $a_{i,j}$ REPRESENTS $P(LA_t = i | LA_{t-1} = j, E_t = happy)$;

(B) THE TRANSITION MATRIX OF LA IN “SURPRISE” EXPRESSION.

(EACH ENTRY $a_{i,j}$ REPRESENTS

$$P(LA_t = i | LA_{t-1} = j, E_t = surprise)$$

expression, the neutral face has high probability to transfer to AU25 (lip apart), and then AU25 is most likely to further transfer to AU25+AU27 (mouth stretch).

In the facial feature level, Figure 4 shows 200 samples drawn from the learned CPDs of the lower-face facial features: $P(LX_t | LA_t)$. (The LX_t in our model is shape difference. For clarity, we show the distribution of LX_t by adding a constant neutral shape: $P(LX_t + C | LA_t)$, where C is a constant neutral shape.) We can see that the local facial feature distributions for different AUs are different. Thus, the AU actually can provide a prior probability of the local shape.

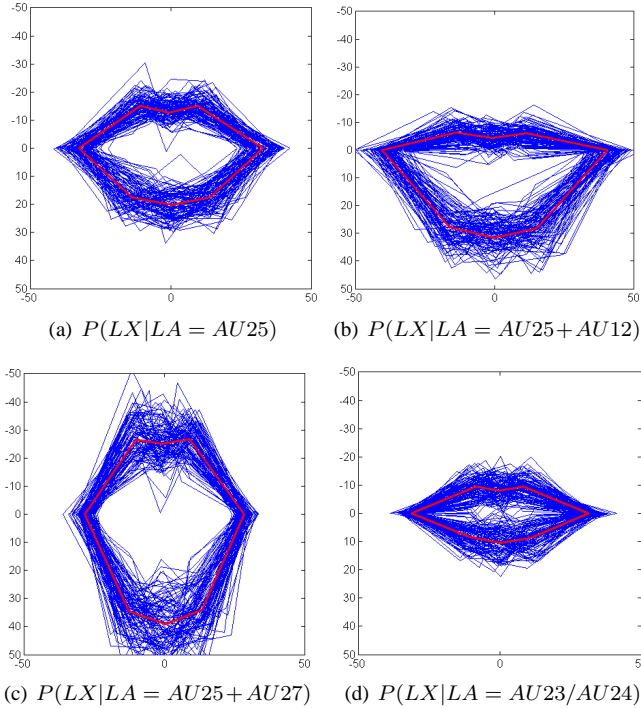


Fig. 4. The CPDs of low-face facial features

IV. MEASUREMENT EXTRACTION

The measurement nodes in our model provide evidence to infer the states of the hidden nodes. Here, we employ various computer vision techniques to acquire various measurements. First, we perform face and eye detection on the face. Given the eye centers, the face region is normalized, and passed through a bank of Gabor filters, and then the classification result for each AU is obtained by the Adaboost classifier similar to [2]. These classification result gives the AU measurements (LAM and UAM) which represent the deformation of the local region.

For the facial feature measurements, we first use the detection method [28] to obtain the facial feature points on the neutral face (the subject is asked to pose neutral expression in the first frame). Then the feature points are tracked using the state-of-the-art facial feature tracker [28], which is based on Gabor wavelet matching and active shape model (ASM). Finally, the shape difference is obtained by comparing the tracked points with the neutral shape. This difference gives the facial feature measurements (LXM and UXM) to our model.

The extracted measurements are not very accurate due to the noise of the image and the limitation of the computer vision technique itself. For example, because the facial feature tracking is based on local search, its performance will decrease during fast movement or video continuity. More specifically, the Gabor wavelet we used for tracking can estimate the feature displacement accurately up to half of the wavelength [29], which is 8 pixel in our experiment. Thus, if the facial feature point movement between two consecutive frames is large than 8 pixel, the tracking error will increase significantly. We expect to improve the robustness by combining all the measurements in our model, and inferring the states of the hidden nodes simultaneously.

V. DBN INFERENCE : SIMULTANEOUS FACIAL FEATURE TRACKING AND EXPRESSION RECOGNITION

Given the DBN model, we want to maximize the posterior probability of the hidden nodes as Equation 3. In our problem, the posterior probability of the expression, AUs and facial features can be computed from the their posterior of the previous frame:

$$\begin{aligned}
& P(E_t, UA_t, LA_t, UX_t, LX_t | UAM_{1:t}, LAM_{1:t}, UXM_{1:t}, LXM_{1:t}) \\
& \propto P(UAM_t | UA_t) P(LAM_t | LA_t) P(UXM_t | UX_t) P(LXM_t | LX_t) \\
& P(UX_t | UA_t) P(LX_t | LA_t) \int_{E_{t-1}, UA_{t-1}, LA_{t-1}} P(E_t | E_{t-1}) \\
& P(UA_t | UA_{t-1}, E_t) P(LA_t | LA_{t-1}, E_t) \\
& P(E_{t-1}, UA_{t-1}, LA_{t-1} | UAM_{1:t-1}, LAM_{1:t-1}, UXM_{1:t-1}, LXM_{1:t-1})
\end{aligned} \tag{6}$$

This filtering problem in DBN can be solved by “Interface algorithm” [18] efficiently.

VI. EXPERIMENTS

A. A Demo Experiment

To demonstrate the effectiveness of our method, we first conduct a demo experiment on a 15-second video sequence with various expressions. To make the experiment more challenging, we capture the video in low-frame rate (6 fps), so that the expression and facial feature point positions can

change significantly in two consecutive frames. We show both the tracking and expression recognition result in detail.

1) *Facial Feature Tracking*: We compare our DBN model with the state-of-the-art facial feature tracker introduced in section IV. The average tracking error (mean square error) for the 14 facial feature points in each frame is shown in Figure 5. The dashed line shows the error of the facial feature tracker, and the solid line is the error of our DBN model. We can see that their performances are very close in most frames, except the sequence after frame 72. We shown the 72nd and 73rd frame in the figure, and the feature points from the baseline system and the DBN model are shown as white and black shapes respectively. We can see that the facial feature tracker fails because it is based on local search, and it cannot track the fast mouth open action in the 73rd frame. The result of the DBN model is better since the strong lower-face AU measurement (AU27+AU25) is detected, and it gives a prior probability of the mouth shape. The average error over the whole sequence is 2.73 pixels for baseline tracker and 2.56 pixels for DBN model. (The average error of eight mouth points is reduced significantly from 2.98 pixels to 2.43 pixels.)

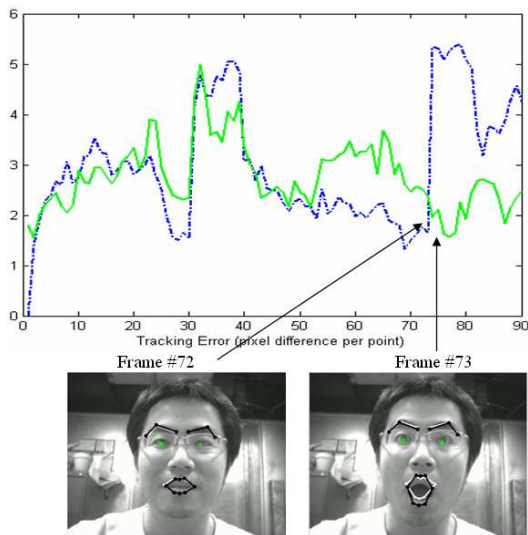


Fig. 5. The tracking error of 14 facial feature points. The dashed line shows the error of the facial feature tracker, and the solid line is the error of the DBN model.

2) *AU and Expression Recognition*: We also compare our model with the Adaboost AU classifier introduced in section IV. For the lower-face AU, the confusion matrices of the Adaboost AU classifier and the DBN model are shown in Table II (Only several large elements of the 8×8 matrix are shown). The Adaboost classifier has the recognition rate of 67%. By combining the mouth shape information from the facial feature level, DBN model increases the recognition rate to 83%. As we can see from the confusion matrices, the DBN model can better distinguish neutral, AU25 and AU25+12.

Meanwhile, the upper face AU recognition rate is also

	Neutral	AU25	AU25+12	AU25+27	AU12	Recognition Rate
Neutral	20	11	2	0	1	57%
AU25	0	4	0	0	1	20%
AU25+12	4	9	20	0	0	61%
AU25+27	0	0	0	16	0	94%
Total Recognition Rate:						67%

(a)

	Neutral	AU25	AU25+12	AU25+27	AU12	Recognition Rate
Neutral	30	1	0	0	4	86%
AU25	1	2	2	0	0	40%
AU25+12	0	3	30	0	0	91%
AU25+27	0	3	1	13	0	76%
Total Recognition Rate:						83%

(b)

TABLE II

THE AU RECOGNITION RESULT. (A) THE CONFUSION MATRIX OF ADABOOST CLASSIFIER. (B) THE CONFUSION MATRIX OF DBN MODEL.

increased from 45% to 69%. Actually, the measurement of the upper face AU is not accurate (only has 45% recognition rate). However, in our learned DBN model, we found that the learned CPD of UAM is close to uniform distribution. Thus, this noisy measurement will not influence other nodes too much. And the states of UA can be inferred from other accurate nodes.

Finally, the expression of each frame is recognized from all the measurements through DBN inference. Figure 6 shows the probability of happy and surprise as a function of time. The expression recognition rate is 81% for this sequence.

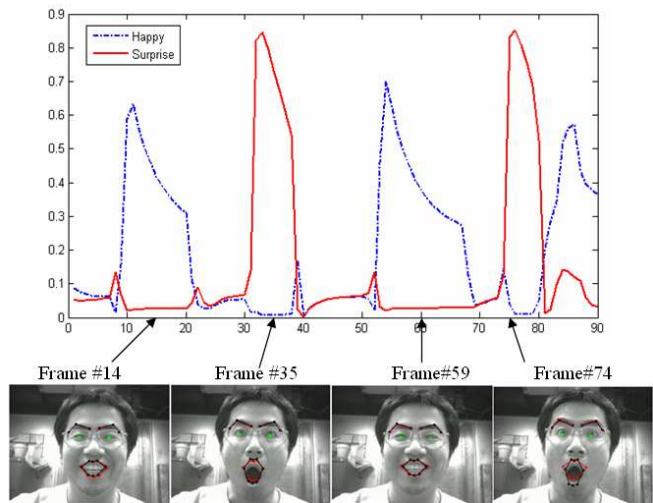


Fig. 6. The probability of “happy” and “surprise” for each frame. The dashed line shows the probability of “happy” and the solid line shows the probability of “surprise”.

B. Result on the Cohn-Kanade Database

The above demo experiment shows how our DBN model simultaneously improves the tracking and recognition performances. Here we conduct experiment on large dataset.

We test our method on Cohn and Kanade’s DFAT-504 database (C-K database) [13], which includes totally 486

	Upper-face AU (UA)	Lower-face AU (LA)	eye points (UX)	mouth points (LX)	all points	Expression
Proposed Method	70.34%	68.88%	2.00 pixels	1.81 pixels	1.89 pixels	60.85%
AU baseline [1]	69.74%	66.41%	–	–	–	–
Feature point tracking baseline [28]	–	–	1.98 pixels	1.99 pixels	1.99 pixels	–

TABLE III

COMPARISON WITH START-OF-THE-ART AU RECOGNITION AND FACIAL FEATURE TRACKING SYSTEMS ON C-K DATABASE. (NOTICE THAT ONLY THE PROPOSED METHOD CAN TRACK DIFFERENT FACIAL ACTIVITIES, I.E. FACIAL FEATURE POINTS, AU AND EXPRESSION, SIMULTANEOUSLY.)

sequences from 97 subjects covering different races, ages, and genders. C-K database has been widely used for evaluating AU recognition system. Kapoor et al. [14] obtained 81.2% recognition rate on 6 AUs of this database using SVM. However, they have to manually label facial feature points for alignment. Bartlett et al. [1] proposed a fully automated AU recognition system, which first uses face detector and Gabor filters to extract features and then uses Adaboost for AU classification. They report overall average recognition rate of 93.6% on 18 AUs, with a true-positive rate (TPR) of 70.39% and false positive rate (FPR) of 2.59%. Yan et al.[26] use a DBN model to capture the spatio-temporal relationships among AUs. This model achieves 93.33% recognition rate on 14 AUs, with a TPR of 86.3% and FPR of 5.5%. By taking the head pose and facial features as the input to DBN model [25], they further improve the TPR to 88.3%.

Some approaches were also proposed to recognize the global expression in the C-K database. Zhao et al. [32] proposed to represent the facial expression using a set of Local Binary Patterns (LBPs), and achieved good recognition accuracy of 96.26% on six basic expressions. Kumano et al. [15] represented the facial expression with the variable-intensity template, which described the intensities of a cloud of points in the vicinity of facial parts. Their method can achieve high recognition rate for a small number of users. But for the experiment on 59 subjects in C-K database, the over-all recognition rate was about 70%.

The previous experiments on C-K database only focus on AU recognition, or global expression recognition, or do them separately. In our hierarchical model, we track the facial features, recognize AU, and recognize global expression simultaneously.

For comparison, we implemented Bartlett et al.’s method [1] as our baseline AU recognition system. In the experiment reported on their website [3], they test on 313 frames of peak AU (i.e., highest magnitude of the target expression) and 313 frames of neutral expression. Using leave-one-subject-out cross validation, the over all recognition rate is 93.6%. In our experiments, we test on 5070 frames from 463 sequences in C-K database, including peak AUs, neutral expressions and weak AUs (low magnitude of target expression). We also use the leave-one-subject-out cross validation to evaluate our baseline system on 11 AUs, as shown in Figure 2. It achieves

91.77% recognition rate, with 80.52% TPR and 5.35% FPR.

The above experiments are focused on classify the binary state (presence/absence) of each AU. However, some AU can be combined to represent different expressions, e.g. AU25+AU27 and AU25+AU12 in Figure 4 represent different mouth expressions. Our method is focused on recognizing these AU combinations. As discussed before, our model classify each frame into one of six upper-face AU combinations and one of eight lower-face AU combinations. For this challenging multi-class classification problem, the baseline system achieves classification rate of 69.74% for upper-face AU and 66.41% for lower-face AU.

To evaluate our facial feature tracking result, we also compare with the state-of-the-art facial feature tracker [28], which achieves average tracking error (mean square error) of 1.98 pixels for eye feature points and 1.98 pixels for mouth feature points.

We summarize the results of our proposed methods and the baseline systems in Table III. We can see that our model can improve both the AU recognition and facial feature tracking results, except the tracking result of eye points. Comparing to eye points, the tracking result of mouth points is improved. Notice that we can observe the same effect in the demo experiment in Section.VI-A.1. The reason is that mouth points undergoes much larger movements than eye points. Therefore the local-based tracker can be improved more with our dynamic model.

Besides more accurate facial feature tracking and AU recognition, our model can recognize 8 global expressions with recognition rate of 60.85%. This result is not as good as the results of state-of-the-art global expression recognition methods.e.g. [15],[32] and [30]. However, notice that we didn’t use any measurement specifically for global expression. Its state is directly inferred from AU and facial feature measurement, and from their relationships.

VII. CONCLUSION

In this paper, we proposed a hierarchical framework for simultaneous facial activity tracking and recognition. By considering the relationships among the facial activities in different levels, the experiments show that it can improve both the tracking and the recognition result.

REFERENCES

- [1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. *Computer Vision and Pattern Recognition, 2005. IEEE Computer Society Conference on*, 2005.
- [2] 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.
- [3] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. <http://mplab.ucsd.edu/grants/project1/research/fully-auto-facs-coding.html>, 2007.
- [4] J. Bazzo and M. Lamar. Recognizing facial actions using gabor wavelets with neutral face average difference. *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, 2004.
- [5] F. Bourel, C. C. Chibelushi, and A. A. Low. Robust facial feature tracking. *Proc. British Machine Vision Conference*, 2000.
- [6] J. M. Buenaposada, E. Munoz, and L. Baumela. Recognising facial expressions in video sequences. *Pattern Analysis and Application*, 11(1):101–116, 2008.
- [7] I. Cohen, N. Sebe, Ashutosh, L. S. C. Garg, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.
- [8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2001.
- [9] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 1995.
- [10] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:974–989, 1999.
- [11] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. *International Journal of Computer Vision (IJCV)*, 76, 2008.
- [12] P. Ekman and W. V. Friesen. *Facial Action Coding System (FACS): Manual*. Consulting Psychologists Press, 1978.
- [13] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, 2000.
- [14] A. Kapoor, Y. Qi, and R. W. Picard. Fully automatic upper facial action recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, (AMFG)*, 2003.
- [15] S. Kumanoo, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision (IJCV)*, 83(2):178–194, 2009.
- [16] J.-J. J. Lien, T. Kanade, J. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous Systems*, 1999.
- [17] K. Murphy. Inference and learning in hybrid bayesian networks. *Report No. UCBCSD-98-990, Computer Science Department, U.C. Berkeley*, 1998.
- [18] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, 2002.
- [19] V. Pavlovic, J. M. Rehg, and J. Maccormick. Learning switching linear models of human motion. In *NIPS*, 2000.
- [20] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
- [21] K. Schwerdt and J. L. Crowley. Robust face tracking using color. In *4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.
- [22] C. Shan, S. Gong, and P. W. MacOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. *Proc. British Machine Vision Conference*, 2006.
- [23] Y.-I. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2001.
- [24] C. Tomasi and T. Kanade. Detection and tracking of point features. *Carnegie Mellon University Technical Report CMU-CS-91-132*, 1991.
- [25] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2010.
- [26] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2007.
- [27] Y. Tong, W. Liao, Z. Xue, and Q. Ji. A unified probabilistic framework for spontaneous facial activity modeling and understanding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [28] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recogn.*, 2007.
- [29] L. Wiskott, J.-M. Fellous, N. Krger, and C. von der Malsburg. Active appearance model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1997.
- [30] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas. Facial expression recognition using encoded dynamic features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [31] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(5), 2005.
- [32] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.
- [33] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 2003.