

# Data-Free Prior Model for Facial Action Unit Recognition

Yongqiang Li, Jixu Chen, Yongping Zhao, and Qiang Ji

**Abstract**—Facial action recognition is concerned with recognizing the local facial motions from image or video. In recent years, besides the development of facial feature extraction techniques and classification techniques, prior models have been introduced to capture the dynamic and semantic relationships among facial action units. Previous works have shown that combining the prior models with the image measurements can yield improved performance in AU recognition. Most of these prior models, however, are learned from data, and their performance hence largely depends on both the quality and quantity of the training data. These data-trained prior models cannot generalize well to new databases, where the learned AU relationships are not present. To alleviate this problem, we propose a knowledge-driven prior model for AU recognition, which is learned exclusively from the generic domain knowledge that governs AU behaviors, and no training data are used. Experimental results show that, with no training data but generic domain knowledge, the proposed knowledge-driven model achieves comparable results to the data-driven model for specific database and significantly outperforms the data-driven models when generalizing to new data set.

**Index Terms**—Facial action units recognition, Bayesian networks, knowledge-driven model



## 1 INTRODUCTION

FACIAL behavior analysis is an important issue in many applications, for example, affective computing, psychological phenomena, agent-human communication. Besides recognizing six basic facial expressions directly, techniques have also been developed to automatically recognize facial action units (AUs). According to the facial action coding system (FACS) developed by Ekman and Friesen [18], AUs represent the muscular activity that produces momentary changes in facial appearance. Although only a small number of distinctive AUs are defined, over 7,000 different AU combinations have been observed so far [27]. Therefore, FACS is demonstrated to be a powerful means for detecting and measuring a large number of facial expressions by virtually observing a small set of muscular actions.

Most current AU recognition techniques are image data driven, and they try to classify each AU or certain AU combinations independently and statically, ignoring the semantic relationships among AUs and the dynamics of AUs. Hence, these approaches cannot always recognize AUs robustly due to the richness, ambiguity, and dynamic nature of facial actions, as well as due to image uncertainty and individual differences. Therefore, prior models are

built to capture the spatial-temporal relationships among AUs. AU recognition can then be performed more robustly by combining the prior model with the image measurements. Hidden Markov models (HMMs) [14], [4], Bayesian network (BN), and dynamic Bayesian network (DBN) [10], [9] are all employed to model the spatial-temporal relationships among AUs and achieved improvement over techniques based on the image observations alone, especially for AUs that are hard to recognize but have strong relationships with other AUs. Furthermore, when the image measurement is not reliable due to either image noise or the inherent deficiencies with image measurement methods, employing a prior model can effectively improve the robustness and the accuracy of the final results.

The use of prior models, however, faces a bottleneck: Learning the model often requires a large amount of reliable and representative training data. Collecting training data (labeling facial actions) often proves to be difficult in real applications, since the effort for training human experts to manually score the AUs is expensive and time-consuming, and the reliability of manually coding AUs is inherently attenuated by the subjectivity of human coder. In addition, despite the best efforts of the database creators, there is always built-in bias in database for computer vision research, such that the model trained on one data set cannot generalize to another data set [3]. Torralba and Efros [3] evaluate the generalization performance of an SVM and off-the-shelf approach [12] for car/person classification/detection task across six databases, which are all collected from internet. The results show that there is a dramatic drop of performance in all tasks and classes when testing on a different test set. For instance, for the car classification task, the average performance obtained when training and testing on the same data set is 53.4 percent, which drops to 27.5 percent when applied to different data sets. For AU recognition problem, prior models learned from data also

• Y. Li and Y. Zhao are with the School of Electrical Engineering and Automation, Harbin Institute of Technology, 92 Xidazhi Street, Harbin 150001, Heilongjiang, China. E-mail: yongqiang.li.hit@gmail.com.

• J. Chen is with the GE Global Research Center, Visualization and Computer Vision Laboratory, One Research Circle, KW-C410, Niskayuna, NY 12308.

• Q. Ji is with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180-3590. E-mail: jiq@rpi.edu.

Manuscript received 6 Sept. 2012; revised 23 Feb. 2013; accepted 11 Mar. 2013; published online 20 Mar. 2013.

Recommended for acceptance by J. Cohn.

For information on obtaining reprints of this article, please send e-mail to: [taffc@computer.org](mailto:taffc@computer.org), and reference IEEECS Log Number TAFCC-2012-09-0071.

Digital Object Identifier no. 10.1109/T-AFFC.2013.5.

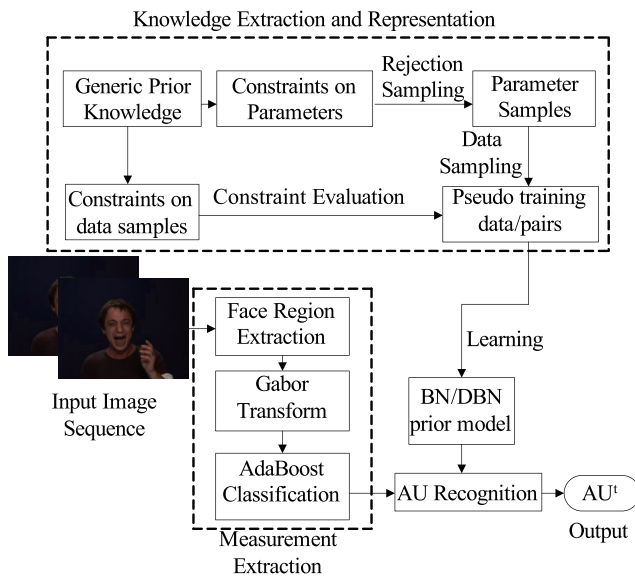


Fig. 1. The flowchart of our AU recognition system.

cannot generalize well to other databases where the relationships of AUs are not represented in the training data. In contrast to the data-driven model, we present a knowledge-driven model, which is exclusively based on domain knowledge, and no training data are used in our approach. Our work contains the following facets:

1. First, we systematically identify and represent the prior knowledge about AUs as constraints on parameters and constraints on data samples.
2. Second, we propose different methods to capture the prior knowledge. Specifically, we introduce an effective sampling method to acquire pseudodata samples and use the distribution of the samples to capture the knowledge.
3. Finally, we propose to learn the prior model from the pseudodata through constrained parameter learning.

Fig. 1 gives the flowchart of our AU recognition system. The system consists of three major components: knowledge extraction and representation, prior model learning, and AU recognition using the prior model and image measurements. The emphasis of this research is on the first two components, where we introduce methods to identify generic AU knowledge, to capture them, and to use them to train the prior model. Given the prior model, AU recognition can be performed by combining the prior model with the image measurements through a probabilistic inference.

## 2 RELATED WORKS

Over the past decades, there has been extensive research in computer vision on facial expression analysis. Current methods in this area can be grouped into two classes: image-driven method and model-based method. In this section, we will present a brief review of the previous works based on these two classes.

### 2.1 Image-Driven Method

Image-driven method for facial action analysis focuses on recognizing facial actions by observing the representative

facial appearance changes. In general, image-driven methods can be divided into two categories: geometric feature-based approach and appearance feature-based approach.

#### 2.1.1 Geometric Feature-Based Approach

Geometric feature-based approaches focus on detecting the location of facial salient points (corners of the eyes, mouth, etc.) [35], [16], and the shapes of the facial components (eyes, mouth, etc.) [2], [41], [42]. The points or shapes are tracked throughout the video, from which features on their relative position, mutual spatial position, speed, acceleration, and so on, are derived. Chang et al. [2] built a probabilistic recognition algorithm based on the manifold subspace of aligned face appearances, which is modeled by 58 facial landmarks. Valstar and Pantic [35] located and tracked a set of facial landmarks and extracted a set of spatial-temporal features from the trajectories, and then, they used a rule-based approach to detect AUs and their temporal segments. Geometric feature-based approaches are more robust to changes in illumination and differences between individuals, but they may fail at some certain AUs, for example, AU15 (Lip Corner Depressor), AU14 (Dimpler), the activation of which involve little displacements of facial fiducial points but changes in skin texture. For extensive survey of facial expression analysis done in the recent years, readers are referred to [30], [40].

#### 2.1.2 Appearance Feature-Based Approach

Facial behavior results in changes of face surface and skin texture. Appearance feature-based approaches try to capture such changes, for example, wrinkles, bulges, furrows. Mahoor et al. [11] transformed 45 AAM-based facial points into Gabor coefficient, and then classify AU combinations using a sparse representation classifier that outperforms SVM and nearest neighbor. Bartlett et al. [21], [15], [22] investigated different features, such as optical flow, explicit feature measurement (i.e., length of wrinkles and degree of eye opening), ICA, and Gabor wavelets, and reported that Gabor wavelets render the best results [22]. Haar features [1], [38] and local binary patterns (LBP) [39] are all well used in expression classification. Tian et al. [16], [17] studied combining the geometric and appearance features together and claimed that the geometric features outperform the appearance-based features, yet using both yields the best result.

Most recently, dynamic appearance descriptors are introduced for activity recognition, which can be seen as an extension of appearance-based approach. Valstar et al. [36] encoded face motion into motion history images (MHI), while Koelstra et al. [32] developed two approaches to model the dynamics and appearances in the face region of an input video: An extended MHI and a method based on nonrigid registration using free-form deformations. Zhao and Pietikaine [37] used volume LBP to recognize dynamic texture and extended to facial image analysis.

### 2.2 Model-Based Method

The common weakness of the image-driven methods is that they tend to recognize each AU or certain AU combination 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. Model-based methods overcome this weakness by making use of the relationships among AUs and recognize various AUs simultaneously. Lien et al. [14] employed a set of HMMs to represent the evolution of facial actions in time. The classification is performed by choosing the AU or AU combination that maximizes the likelihood of the extracted facial features generated by the associated HMM. Valstar and Pantic [4] used a combination of SVMs and HMMs and outperformed the SVM method for almost every AU by considering the temporal evolution of facial action. Both methods exploit the temporal dependencies among AUs. They, however, fail to exploit the spatial dependencies among AUs. To remedy this problem, Tong et al. [10], [9] employed a DBN to systematically model the spatiotemporal relationships among AUs and achieved a marked improvement over the image observation, especially for AUs that are hard to recognize but have strong relationships with other AUs. The use of prior model can effectively handle the noisy image observation, but the data-driven models suffer the following drawbacks: First, training the data-driven model needs a large amount of annotated and representative data, which sometimes proves to be hard to achieve for AU recognition problem; Second, data-driven prior model depends on specific database [10], and cannot generalize well to other databases. A separate DBN model is, therefore, needed for each data set.

Recently, to address this issue, researchers in machine learning try to incorporate domain knowledge into model learning process to reduce the dependence on training data. Most of these approaches incorporate qualitative prior knowledge, for example, constraints on parameters, into the parameters learning process by formulating the learning as a constrained optimization problem [5], [8], [7]. While effective, often with a closed solution, the knowledge constraints used by these methods are limited to simple linear constraints on parameters. Liao and Ji [8] included more complex constraints with an iterative optimization procedure. Campos and Ji [13] proposed a method that allows both hard constraints and soft constraints. Mao and Lebanon [6] used soft Bayesian prior to regulate the maximum likelihood (ML) score and introduce the concept of model uncertainty with a maximum a posterior estimation. There are two main limitations with these approaches: First, most of these approaches do not explain the source of the constraints, and the domain knowledge they use is limited to a few simple qualitative constraints; Second, the qualitative constraints in previous works are used as supplementary information to data. During training, data are still used.

### 2.3 Outstanding Features of Our Approach

In this paper, we propose a knowledge-driven method to learn a prior AU model from different types of qualitative knowledge. Compared to previous works, the proposed method has the following features:

1. First, in contrast to the data-driven model, our knowledge-driven model is totally learned from the generic domain knowledge, and no training data

are used. Therefore, our model has no dependence on the data and can generalize well to different data sets. This is practically significant since acquiring the annotated training data is an expensive, subjective, and time-consuming process.

2. Second, although some methods have been proposed to incorporate prior knowledge into model learning, they, however, are limited to some simple parameter constraints. And these methods still need training data. In our method, we impose various prior knowledge into our AU prior model without using any training data.
3. Third, we introduce a unified Markov chain Monte Carlo (MCMC) sampling method to simultaneously incorporate these knowledge into the DBN model learning by first converting the generic knowledge into synthetic data, and then using the conventional learning method to train the prior model from the synthetic data. The new learning method allows simultaneously incorporating different types knowledge into the prior model in a principled manner.

In the remainder of this paper, we discuss our knowledge-driven method in detail. We first present the definition of the generic prior knowledge that we employ (Section 3). Then, a knowledge-driven method to learn a AU prior model is proposed (Section 4). We demonstrate the effectiveness of our method on two databases and compare with the data-driven model in Section 5.

## 3 GENERIC KNOWLEDGE ON FACIAL ACTIONS

In this section, we introduce the generic knowledge on facial actions. They can be expressed as qualitative constraints on individual AUs (Section 3.2), on group AUs (Section 3.3), and on AU dynamics (Section 3.4). As discussed below, the generic knowledge we used is primarily from the study of the FACS, the consultation with psychologists, an empirical analysis of facial anatomy, and the previous studies. When extracting knowledge from the databases, we strive to extract the general knowledge that is applicable to all databases. Such knowledge only supplements to the knowledge derived from the theories.

### 3.1 Causal Influence among AUs

According to FACS, there are a total of 33 exclusive facial action descriptors, 30 of which are anatomically related to the contraction of a specific set of facial muscles, which generally lie from skull to skin and are innervated by facial nerve. Unlike other skeletal muscles that attach to bones, facial muscles attach to each other or to the skin. Fig. 2 shows facial muscles anatomy. Some facial muscles, their actions, and the corresponding AUs are summarized in Table 1.

Facial actions are related to each other both spatially and dynamically to form a coherent and consistent facial expression [29]. Through the study of the FACS [29], an empirical analysis of facial anatomy and the consultation with psychologists, we derive some constraints that govern the motion of the facial actions. For example, AU2 (outer brow raiser) and AU1 (inner brow raiser) are both related to the muscle group of *Occipito frontalis*, as shown in Table 1,

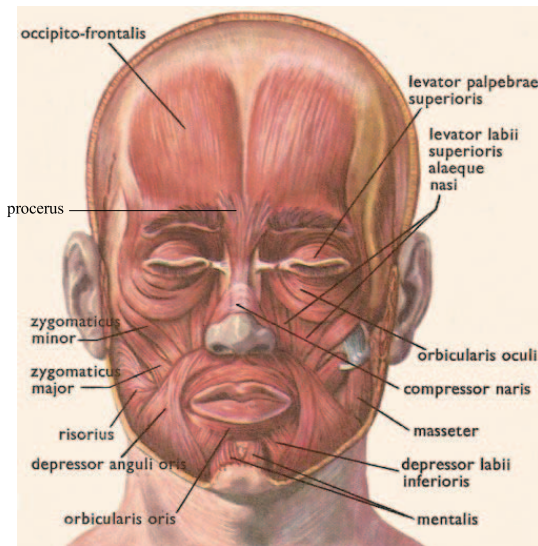


Fig. 2. Facial muscle anatomy. There are a total of 17 facial muscles controlling different facial actions (adapted from [44]).

which is in the scalp and forehead that raises the eyebrows. The contraction of the lateral part of this muscle group produces AU2, while the contraction of the medial (or central) portion of this muscle group produces AU1. Hence, "AU2 is a difficult movement for most people to make voluntarily without adding AU1" as described in FACS [29], which means the appearance of AU2 increases the probability of the occurrence of AU1, and we call this a positive influence from AU2 to AU1. To represent this qualitative influence graphically, we link AU2 node to AU1 node with a "+" sign to denote positive influence as shown in Fig. 3.

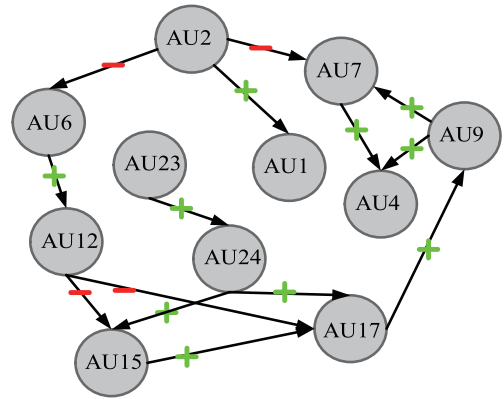
















Fig. 3. AU causal influence network.

On the other hand, there are some other AUs whose appearance will decrease the chance of the occurrence of another AU. For example, when AU12 occurs (lip corner puller), which is produced by the muscle group of *Zygomaticus Major*, it will decrease the chance of the occurrence of AU15 (lip corner depressor), which is produced by the muscle group of *depressor anguli oris*. We call this a negative influence from AU12 to AU15 and add a link with "-" sign from AU12 to AU15. There are many such empirical constraints, for example, mouth stretch increases the chance of lips apart and decreases the chance of cheek raiser and lip presser; cheek raiser and lid compressor increases the chance of lip corner puller; outer brow raiser increases the chance of inner brow raiser and decreases the chance of nose wrinkler; lip tightener increases the chance of lip presser; lip presser increases

TABLE 1  
Muscles, Actions, and Corresponding AUs

| Facial Muscles                        | Actions   | AUs   |
|---------------------------------------|---|---|
| Occipito Frontalis                    | Draw scalp forward and raises eyebrows              | AU1  AU2    |
| Procerus                              | Pull the glabella down                              | AU4    |
| Orbicularis Oculi                     | Spread tears across cornea and close eyelid tightly | AU6  AU7    |
| Levator Labii Superioris Alaeque Nasi | Raise upper lip and wrinkles the nose               | AU9  AU10   |
| Zygomaticus Major and Minor           | Draw angle of mouth upward                          | AU11  AU12  |
| Depressor Anguli Oris                 | Draw angle of mouth downward                        | AU15   |
| Depressor Labii Inferioris            | Lowers lower lip                                    | AU16   |
| Mentalis                              | Draws chin up                                       | AU17   |
| Orbicularis Oris                      | Levator/depressor of lip and angle of mouth         | AU23  AU24  |

the chance of lip corner depressor and chin raiser. Through the way we analyzed above, we construct a causal influence network to represent these qualitative influence constraints, as shown in Fig. 3, where every link between two AU nodes has a sign to capture either the positive or negative influence between two AUs, with a positive sign denoting positive influence and a negative sign denoting negative influence.

### 3.2 Constraints on Individual AUs

Given the causal influence network, we can extract two types of constraints on an  $AU_i$ , depending on the number of AUs that influence  $AU_i$ . If  $AU_i$  is either positively or negatively influenced by only one AU (e.g.,  $AU_6$  in Fig. 3) and let that AU be  $AU_j$ , we can then construct the following constraints:

$$P(AU_i = 1 | AU_j = 1) > P(AU_i = 1 | AU_j = 0) \quad (1)$$

if  $AU_j$  positively influences  $AU_i$

$$P(AU_i = 1 | AU_j = 1) < P(AU_i = 1 | AU_j = 0) \quad (2)$$

if  $AU_j$  negatively influences  $AU_i$ .

If, on the other hand,  $AU_i$  is influenced by multiple  $AU_s$  (e.g.,  $AU_4$  in Fig. 3), assuming all influences are the same (e.g., all are positive or all are negative) and denote all positive influencing AUs as  $AU^P$  and all negative influencing AUs as  $AU^N$ , we can construct the following constraints:

$$P(AU_i = 1 | AU^P = 1) > P(AU_i = 1 | AU^P \neq 1) \quad (3)$$

$$P(AU_i = 1 | AU^N = 1) < P(AU_i = 1 | AU^N \neq 1), \quad (4)$$

where  $AU^P = 1$  (or  $AU^N = 1$ ) means the values of all elements of  $AU^P$  ( $AU^N$ ) are positive, while  $AU^P \neq 1$  (and  $AU^N \neq 1$ ) means that the values of some elements of  $AU^P$  ( $AU^N$ ) are not equal to 1, i.e., 0.

Finally, if  $AU_i$  (e.g.,  $AU_{15}$  or  $AU_{17}$ ) is influenced by a combination of both positive AUs ( $AU^P$ ) and negative AUs ( $AU^N$ ), we can construct the following constraint:

$$\begin{aligned} &P(AU_i = 1 | AU^P = 1, AU^N = 0) \\ &> \left\{ \begin{array}{l} P(AU_i = 1 | AU^P = 0, AU^N = 0) \\ P(AU_i = 1 | AU^P = 1, AU^N = 1) \end{array} \right\} \quad (5) \\ &> P(AU_i = 1 | AU^P = 0, AU^N = 1). \end{aligned}$$

Besides casual qualitative influence among AUs, there is also distribution constraint on some AUs. In spontaneous cases, some AUs (e.g.,  $AU_2$ ) less likely occur. This means the probability of some AUs in specific states is higher than these AUs in other states. This type of knowledge can be defined by a single distribution constraint. Let  $AU_i$  be such an AU, we then have

$$P(AU_i = 1) < P(AU_i = 0), \quad (6)$$

where 1 means AU presence and 0 means AU absence.

### 3.3 Constraints on Group AUs

Activating the AUs produces significant changes in the shape of facial component. For example, activating  $AU_{27}$

TABLE 2  
AU Combinations with Low Probability to Occur in Spontaneous Facial Expressions

| Eyebrow movement group            | Mouth movement group                       |
|-----------------------------------|--|
| $P(AU_1 = 0, AU_2 = 1, AU_4 = 1)$ | $P(AU_{12} = 1, AU_{15} = 0, AU_{17} = 1)$ |
|                                   | $P(AU_{12} = 1, AU_{15} = 1, AU_{17} = 0)$ |
|                                   | $P(AU_{12} = 1, AU_{15} = 1, AU_{17} = 1)$ |

results in a widely open mouth; and activating  $AU_4$  makes the eyebrow lower and pushed together. As a result, the corresponding local facial component movements are also controlled by the AUs. We divide the AUs we are going to recognize into three groups based on facial component:

1. *Eyebrow group.*  $AU_1$ ,  $AU_2$ , and  $AU_4$ , controlling eyebrow movements.
2. *Eyelids group.*  $AU_6$  and  $AU_7$ , controlling eyelids movements.
3. *Mouth group.*  $AU_{12}$ ,  $AU_{15}$ , and  $AU_{17}$ , controlling mouth movements.

In each group, we analyze the co-occurrence/coabsence of the corresponding AUs based on their underlying muscles and then derive corresponding probabilistic constraints. For example, three AUs ( $AU_{12}$ ,  $AU_{15}$ ,  $AU_{17}$ ) control mouth movement, and through the empirical analysis of facial anatomy and FACS [29], we found that,  $AU_{15}$  and  $AU_{17}$  rarely occur with  $AU_{12}$  because of the facial muscular constraints (as analyzed in Section 3.1). Some previous studies, i.e., [10], also provide similar supplemental evidences. Based on this understanding, we list the AU combinations that have a low probability to occur for each group in Table 2. The low probability for a combination in Table 2 can be expressed as a constraint that the probability of a AU combination is lower than the probability of any other combination of the same group of AUs. For example, for the eyebrow group, we have

$$\begin{aligned} &P(AU_1 = 0, AU_2 = 1, AU_4 = 1) \\ &< P(AU_1 = 1, AU_2 = 1, AU_4 = 1). \end{aligned} \quad (7)$$

( $AU_1 = 1, AU_2 = 1, AU_4 = 1$ ) is one possible configuration of the three AUs. There are a total of seven such configurations, hence producing seven such constraints.

### 3.4 Constraints on AU Dynamics

Besides the static constraints, there are also dynamic constraints that restrict the temporal evolutions among AUs. In this work, we consider the following dynamic constraints:

1. *AU level dynamic constraint.* We assume that each individual AU varies smoothly in a spontaneous expression. We can then model the relationship between the state of AU in next time step  $AU^{t+1}$  and its state in current time step  $AU^t$  as follows:

$$P(AU^{t+1} = s | AU^t = s) > P(AU^{t+1} \neq s | AU^t = s), \quad (8)$$

where  $s$  represents a binary state of 0 or 1.

TABLE 3

AU Combinations between Two Consecutive Time Steps with Low Probability to Occur in Spontaneous Facial Expressions

| Low probability to occur                           |
|--|
| $P(AU_6^t = 1, AU_{12}^{t-1} = 0, AU_6^{t-1} = 0)$ |
| $P(AU_6^t = 0, AU_{12}^{t-1} = 1, AU_6^{t-1} = 1)$ |
| $P(AU_1^t = 1, AU_2^{t-1} = 0, AU_1^{t-1} = 0)$    |
| $P(AU_1^t = 0, AU_2^{t-1} = 1, AU_1^{t-1} = 1)$    |

2. *Expression level dynamic constraint.* In spontaneous facial behaviors, some AUs usually occur together to express certain emotion. Furthermore, the multiple AUs involved may not undergo the same development simultaneously; instead, they often proceed in sequence as the intensity of facial expression varies. For example, Schmidt et al. [23] found that certain AUs usually closely followed the appearance of AU12 in smile expression. For 88 percent of the smile data they collect, the appearance of AU12 was either simultaneously with or closely followed by one or more associated AUs, and for these smiles with multiple AUs, AU6 was the first AU to follow AU12 in 47 percent. Messinger et al. [45] also show that AU6 may follow AU12 (smile) or AU20 (cry) to act as an enhancer to enhance the emotion. This means that certain AU in next time step may be affected by other AUs in the current time step. Analysis of other expressions results in a similar conclusion. For example, in “brow raise” expression, AU1 usually follows the AU2 to enhance the expression. Similar findings are found in [10]. Based on this understanding, we can obtain the expression level dynamic constraint: Some AUs have strong dynamic dependencies, while other AUs have little or no dynamic dependencies. AUs that are strongly dependent on each other dynamically include AU12, AU6, and AU2, AU1. For example, AU12 often precedes AU6, while AU2 is often followed by AU1. The AU combinations between two consecutive time steps that have low probability to occur in spontaneous expressions are listed in Table 3.

## 4 KNOWLEDGE-DRIVEN MODEL LEARNING

We propose a knowledge-driven method to learn a prior AU model based on the above constraints. The learning process is composed of two steps. The first step is to produce the feasible model parameter samples that satisfy the constraints. This is then followed by converting the feasible model parameter samples into pseudodata (Section 4.2). Finally, the prior AU model is learned from the pseudodata (Section 4.3).

### 4.1 BN as the AU Prior Model

#### 4.1.1 A BN Model for AU Recognition

Following the work in [10], [9], we propose to use the BN as the prior model to capture the AU knowledge and to perform AU recognition. The prior model probabilistically encodes the soft and probabilistic constraints to capture the

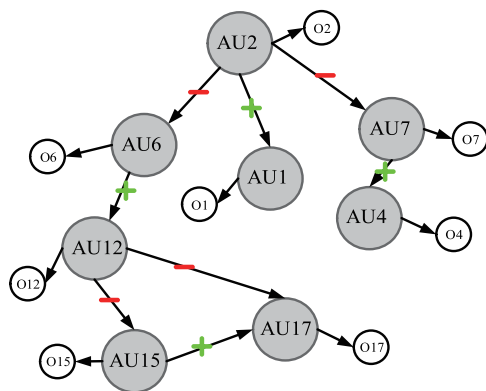


Fig. 4. A BN AU prior model.

AU occurrence frequency. While correct for the majority of the samples, these soft constraints may not be consistent with every sample or certain expressions. For example, AU12 decreases the probability of AU15 and AU17, but for certain expressions, for example, smile control, AU12 may occur frequently with AU15 and AU17 [48], [49]. While this is a weakness of the proposed prior model, it is in fact the case with any prior models since the prior models typically improve the overall performance, but cannot guarantee to be correct for every case.

A BN is a directed acyclic graph that represents a joint probability distribution among a set of random variables. Based on consultation with the domain expert and on the work in [10], we construct a BN as shown in Fig. 4 to capture the dependencies among facial AUs. In this BN model, big-shaded nodes represent AUs, each of which has two states (presence 1 and absence 0). The AU nodes are hidden, and their true states are unknown. The small nodes represent the corresponding image measurement of the hidden AU nodes. A BN can be uniquely determined by a structure and a set of parameters. The parameters of a BN consist of the conditional probability distribution (CPD) for each node, given its parents. The AU constraints discussed in Section 3.2 can be translated into constraints on the BN parameters. For example, parameters for AU4 node are two conditional probabilities,  $P(AU4 | AU7 = 1)$  and  $P(AU4 | AU7 = 0)$ . Since AU7 has positive influence on AU4, according to Fig. 4, we can get  $P(AU4 | AU7 = 1) > P(AU4 | AU7 = 0)$  as per (1), and we call this a monotonicity constraint on our BN model parameters. There are such constraints for each node except for the root node AU2, which follows the distribution constraint as per (6), that can be expressed as  $P(AU2 = 1) < P(AU2 = 0)$ . The constraints on the BN model parameters are summarized in Table 4.

#### 4.1.2 A DBN Model for AU Recognition

The BN model we constructed can just model the static relationships among AUs. To capture the dynamic dependencies, we extend our model to DBN, which models the temporal evolution of a set of random variables  $X$  over time. A DBN can be defined by a pair of BNs ( $B_0, B_-$ ): 1) the static network  $B_0$ , as shown in Fig. 5a, captures the static distribution over all variables  $X^0$  in the initial time frame; and 2) the transition network  $B_-$ , as shown in Fig. 5b, specifies the transition probability  $P(X^{t+1} | X^t)$  for all  $t$  in

TABLE 4  
Constraints on BN Model Parameters

| AU node | Constraints  |
|---------|--|
| AU2     | $P(AU2 = 1) < P(AU2 = 0)$  |
| AU1     | $P(AU1 AU2 = 1) > P(AU1 AU2 = 0)$  |
| AU4     | $P(AU4 AU7 = 1) > P(AU4 AU7 = 0)$  |
| AU6     | $P(AU6 AU2 = 1) < P(AU6 AU2 = 0)$  |
| AU12    | $P(AU12 AU6 = 1) > P(AU12 AU6 = 0)$  |
| AU15    | $P(AU15 AU12 = 1) < P(AU15 AU12 = 0)$  |
| AU7     | $P(AU7 AU2 = 1) < P(AU7 AU2 = 0)$  |
| AU17    | $P(AU17 AU15 = 0, AU12 = 1) < \left\{ \begin{array}{l} P(AU17 AU15 = 0, AU12 = 0) \\ P(AU17 AU15 = 1, AU12 = 1) \end{array} \right\} < P(AU17 AU15 = 1, AU12 = 0)$ |

finite time slices  $T$ . Given a DBN model, the joint probability over all variables  $X^0, \dots, X^T$  can be factorized by “unrolling” the DBN into an extended static BN, as shown in Fig. 5c, whose joint probability is computed as follows:

$$P(x^0, \dots, x^T) = P_{B_0}(x^0) \prod_{t=0}^{T-1} P_{B_-}(x^{t+1} | x^t), \quad (9)$$

where  $x^t$  represents the sets of values taken by the random variables  $X$  at time  $t$ ,  $P_{B_0}(x^0)$  captures the joint probability of all variables in the static BN  $B_0$ , and  $P_{B_-}(x^{t+1} | x^t)$  represents the transition probability that can be decomposed as

$$P_{B_-}(x^{t+1} | x^t) = \prod_{i=1}^N P_{B_-}(x_i^{t+1} | pa(x_i^{t+1})), \quad (10)$$

where  $pa(x_i^{t+1})$  represents the parent configuration of variable  $x_i^{t+1}$  in the transition network  $B_-$ .

In this work, besides the dynamics within a single AU, which depicts how a single  $AU_i$  develops over time, we also consider the dynamics among different AUs. As discussed in Section 3.4, there are expression level dynamic constraints between two consecutive time steps, so we manually set two dynamic links between different AUs, which are from AU12 at time  $t-1$  to AU6 at time  $t$  and from AU2 at time  $t-1$  to AU1 at time  $t$ , respectively, to capture such constraints. Finally, the DBN structure as shown in Fig. 6 is used to capture the spatial-temporal relationships among AUs. The temporal links, i.e., the self-pointed arrows and the dynamic links between AUs at two

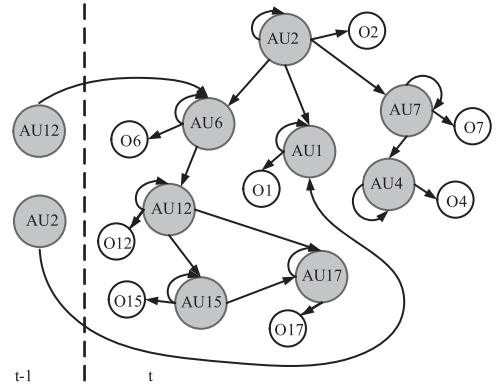


Fig. 6. The DBN for AU modeling. The self-arrow at each AU node indicates the temporal relationship of a single AU from the previous time step to the current time step. The arrow from  $AU_i$  at time  $t-1$  to  $AU_j$  ( $j \neq i$ ) at time  $t$  indicates the temporal relationship between different AUs. The small circle indicates the measurement for each AU.

time slices are used to impose the dynamic AU constraints discussed in Section 3.4, i.e., the temporal smoothness constraint and the dynamic dependence constraint, respectively. In the following section, we discuss the method to learn the BN/DBN parameters from knowledge constraints.

## 4.2 Generating Parameter Samples and Pseudodata

In this section, we first introduce a sampling approach to efficiently acquire the BN AU model parameter samples (Section 4.2.1). Then, based on the parameter samples, we generate pseudotraining data and pseudodata pairs (Section 4.2.2).

### 4.2.1 Generating BN Parameter Samples

Based on the AU constraints discussed above, we propose to generate parameter sample, i.e., a vector of CPDs for all nodes. To effectively generate many instances satisfying the parameter constraints as listed in Table 4, we use rejection sampling method [20], which consists of two steps: First, generate samples from a proposal distribution and then reject the samples inconsistent with constraints. The second step is simply checking the sample with each constraint. The first step is more difficult because we need to generate samples in a high-dimensional space. To explore the space efficiently, we propose the following sampling method.

The basic idea is to generate more samples from the current “unexplored” region, so that the whole space can be explored more efficiently. Specifically, we define the proposal distribution of the  $l$ th parameter sample  $\theta^l$

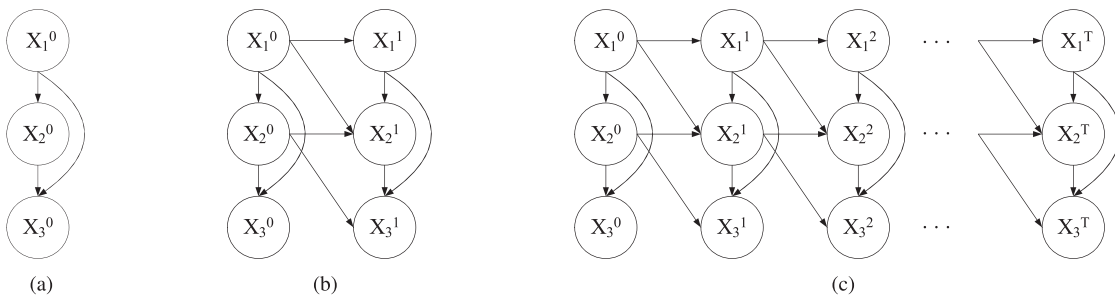


Fig. 5. A pair of (a) static network  $B_0$  and (b) transition network  $B_-$  defines the dynamic dependencies for three random variables  $X_1$ ,  $X_2$ , and  $X_3$ . (c) The corresponding “unrolled” DBN for  $T+1$  time slices.

conditioned on the previous instances:  $p(\theta^l | \theta^{l-1}, \dots, \theta^1)$ . This probability is higher when the sample is far from previous instances.

Given the previous instances, we first define a kernel density function with Gaussian kernel:

$$q(\theta^l | \theta^{l-1}, \dots, \theta^1) = \frac{1}{l-1} \sum_{j=1}^{l-1} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\theta^l - \theta^{(j)}\|^2}{2\sigma^2}\right\}, \quad (11)$$

where  $D$  is the dimension of the sample (i.e., number of model parameter),  $\sigma$  represents the standard deviation (SD). This function has high probability in the region close to previous samples. Since we need to explore the regions that have not been explored, our proposal distribution is defined as follows:

$$p(\theta^l | \theta^{l-1}, \dots, \theta^1) \propto 1/(2\pi\sigma^2)^{D/2} - q(\theta^l | \theta^{l-1}, \dots, \theta^1). \quad (12)$$

$1/(2\pi\sigma^2)^{D/2}$  is the largest possible value of  $q(\theta^l | \theta^{l-1}, \dots, \theta^1)$ . Now, the problem is how to generate a new sample  $\theta$  according to this proposal distribution. Considering the constraints, we use the rejection sampling method as follows:

1. We first generate each element of a sample  $\theta^l$  from a uniform distribution.
2. If  $l = 1$ , this sample is always accepted; otherwise, this sample is accepted with a probability  $\frac{p(\theta^l | \theta^{l-1}, \dots, \theta^1)}{1/(2\pi\sigma^2)^{D/2}}$ . This can be easily implemented as subroutine:
  - a. Generate a number  $u$  from the uniform distribution over  $[0, 1/(2\pi\sigma^2)^{D/2}]$ ;
  - b. if  $u < p(\theta^l | \theta^{l-1}, \dots, \theta^1)$ ,  $\theta^l$  is accepted; otherwise, it is rejected.
3. If  $\theta^l$  is rejected, go back to Step 1 to generate another sample, until the new sample is accepted.
4. Check the new sample (CPD) with the parameter constraints as listed in Table 4, if the sample satisfy all the constraints, add the new sample to the sample set  $\theta^l \rightarrow C$ , otherwise reject this sample and go back to Step 1.
5. If the sample set size  $|C|$  is smaller than  $L$ , then go back to Step 1.

We can see that this algorithm includes two rejection steps. Each sample is first tested by the proposal distribution to make it far from previous instances. Then, the sample is tested by the parameter constraints. Finally, we can get a concise sample set satisfying the parameter constraints. Given the parameter samples, we can simply find their mean and use the mean as the parameters of the prior model. But doing so will lose modeling accuracy since the sampled parameters do not follow a Gaussian distribution. We instead propose to generate pseudodata samples from each parameter sample, which are further evaluated by the constraints on data samples. The valid pseudosamples are collectively used to train the prior model as detailed below.

#### 4.2.2 Generating Pseudodata

Based on the BN parameter sample  $\theta_i$  we generated above, now we generate pseudodata  $D_j$ , which is a vector representing the AUs states in one time instant. Each BN

parameter sample  $\theta_i$  and the BN structure together define a joint probability distribution and represent the constraints. We drew 500 samples, for example,  $D_j, j = 1, \dots, 500$ , from each  $\theta_i, i = 1, \dots, k$ . Then, we combine all the samples together as the pseudotraining data set. We use AU group constraints to evaluate the data samples as follows:

We first generate a data sample  $D_j$  from the joint probability defined by  $\theta_i$ , check  $D_j$  with AU group constraints as listed in Table 2. If there is no such instance,  $D_j$  is accepted; otherwise,  $D_j$  is accepted with probability  $p$  (we set it 0.1 in this work).

The pseudodata we generate above can only represent static constraints. To incorporate the dynamic constraints, we propose to generate pseudodata pairs  $(D^{t+1}, D^t)$  that include data of both the current and the next time step. This dynamic sampling procedure is summarized as follows:

1. Sample the current time step data  $D^t$  using the above method.
2. Given  $D^t$ , we generate the next time step data  $D^{t+1}$  according to the AU level dynamic constraint. Since this constraint is imposed on each AU separately, we sampled each element of  $D^{t+1}$  independently satisfying the AU level dynamic constraint.
3. Check  $D^{t+1}$  with AU group constraints and check pseudo data pair  $(D^{t+1}, D^t)$  with the expression level dynamic constraint as listed in Table 3, respectively. If this pair is infeasible, then reject it and go back to Step 1.

#### 4.3 Learning BN/DBN Parameters from Constraints

Given the BN/DBN structure, now we focus on learning the parameters from pseudotraining data to infer each AU. We first introduce the BN parameter learning method and then extend it to DBN. Learning the parameters in a BN is to find the most probable values  $\hat{\theta}$  for  $\theta$  that can best explain the generated pseudotraining data. Let  $\theta_{ijk}$  indicates a probability parameter for a BN,

$$\theta_{ijk} = p(x_i^k | pa^j(X_i)), \quad (13)$$

where  $i$  ranges over all the variables (nodes in the BN),  $j$  ranges over all the possible parent instantiations for variable  $X_i$ , and  $k$  ranges over all the instantiations for  $X_i$  itself. Therefore,  $x_i^k$  represents the  $k$ th state of variable  $X_i$ , and  $pa^j(X_i)$  is the  $j$ th configuration of the parent nodes of  $X_i$ . In this work, the "fitness" of parameters  $\theta$  and training data  $D$  is quantified by the log-likelihood function  $\log(p(D | \theta))$ , denoted as  $L_D(\theta)$ . Assuming the pseudodata we generated are independent, based on the conditional independence assumptions in BNs, we have the log-likelihood function in

$$L_D(\theta) = \log \prod_{i=1}^n \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}}, \quad (14)$$

where  $n_{ijk}$  is the count for the case that node  $X_i$  has the state  $k$ , with the state configuration  $j$  for its parent nodes;  $n$  is the number of variables (nodes) in the BN;  $q_i$  is the number of parent configurations of  $X_i$  node; and  $r_i$  is the number of instantiations of  $X_i$ .



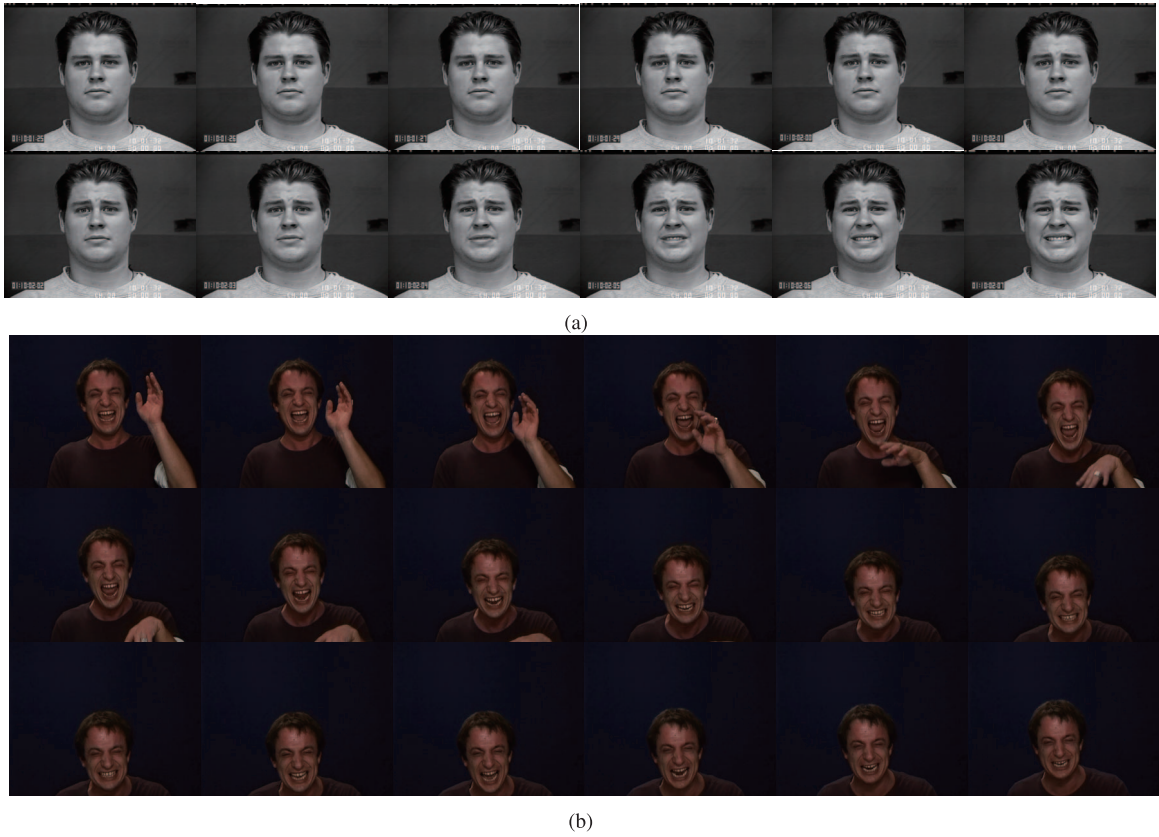


Fig. 7. (a) Some example images from CK database (adapted from [19]). (b) Some example images from FERA database.

Since we have got the complete pseudotraining data, an ML estimation method can be described as a constrained optimization problem, i.e., maximize (15), subject to  $n$  equality constraints (16):

$$\text{Max} \quad L_D(\theta) \quad (15)$$

$$\text{S.T.} \quad g_{ij}(\theta) = \sum_{k=1}^{r_i} \theta_{ijk} - 1 = 0, \quad (16)$$

where  $g_{ij}$  imposes the constraint that the parameters of each node sums to 1 over all the states of that node,  $1 \leq i \leq n$  and  $1 \leq j \leq q_i$ . Solving the above equations, we can get

$$\theta_{ijk} = \frac{n_{ijk}}{\sum_k n_{ijk}}.$$

Since a DBN can be seen as a pair of BN ( $B_0, B_{\rightarrow}$ ) and the static network  $B_0$  is the same as the BN we learned above, we only need to learn the transition network  $B_{\rightarrow}$ . In implementation, we consider each pseudodata pair as one data sample for the transition network  $B_{\rightarrow}$ , then we use the same learning method above to learn the parameters of the transition network  $B_{\rightarrow}$ . Then, we combine the  $B_0$  and  $B_{\rightarrow}$  together as the DBN model for AU recognition in the following section.

## 5 EXPERIMENTS

### 5.1 Facial Expression Database

The proposed knowledge-driven model is tested on FACS labeled images from two databases. The first database is the

Cohn-Kanade DFAT-504 (C-K) database [24], which consists of more than 100 subjects covering different races, ages, and genders. To extract the temporal relationships, the C-K database is coded into AU labels frame by frame in this work.

Furthermore, the FG 2011 facial expression recognition and analysis challenge (FERA) database [28] is employed to evaluate the generalization ability of our knowledge-driven model. FERA database is a subset of the GEMEP corpus [47], in which the subjects are all professional actors and are coached by a professional director. The main differences between Cohn-Kanada database and FERA database are as follows: 1) The image sequences on FERA database contain a complete temporal evolution of expression while that on C-K database only reflect the evolution of the expression starting from a neutral state and ending at the apex, but without the relaxing period, 2) subjects on FERA database are asked to perform spontaneous expression with natural head movements, while subjects on C-K database only perform simple AU combinations in frontal view face. Examples from these two databases are shown in Fig. 7.

### 5.2 AU Measurement Extraction

When we estimate the AU state from the image, this prior model is combined with image measurements to estimate the posterior probability of AUs. In this work, we employ Gabor features and an AdaBoost classifier for AU measurements extraction. For each image, we first detect the eyes through a boosted eye detector [43]. Then, the image is normalized into  $64 \times 64$  subimage based on the eye positions. A set of six orientations and five scales Gabor

filters are applied, and a  $6 \times 5 \times 64 \times 64 = 122,880$  dimension feature vector is obtained for each image. Given the image features, the AdaBoost classifier is then employed to obtain the measurement for each AU. Through the training process, the weights of the wrongly classified examples are increased in each iteration, and AdaBoost forces the classifier to focus on the most difficult samples in the training set. And, thus, it results in an effective classifier. In this work, the final classifier utilizes around 200 Gabor features for each AU. Based on the image measurement  $e_i$  and ground truth  $AU_i$ , we then train a likelihood function that is a conditional probability of the AU measurement given the actual AU values,  $P(e_i | AU_i)$ . Note that we still need training data to train the AU measurement method. But such data are not used to train the prior model since prior model trained using such data cannot generalize well to a different data set as shown in our experiments. Moreover, training a prior model typically needs much more data than training a measurement model.

### 5.3 AU Recognition through BN/DBN Inference

In the above sections, we have learned the BN/DBN model to represent the prior probability of AUs. Once the image measurements are obtained, we can use them as the evidence to estimate the true state of AUs through BN/DBN model inference. Let  $AU_i$  indicate the  $i$ th AU node, and  $e_i$  be the corresponding measurement. In BN inference, the posterior probability of AUs can be estimated by combining the likelihood from measurement with the prior probability of AUs:

$$p(AU_1, \dots, AU_N | e_1, \dots, e_N) \propto \prod_{i=1}^N p(e_i | AU_i) \prod_{i=1}^N p(AU_i | Pa(AU_i)). \quad (17)$$

The first term is the likelihood term. The second term is the product of the conditional probabilities of each AU node  $AU_i$  given its parents  $Pa(AU_i)$ , which are BN model parameters that have been learned. In practice, the posterior probability of each AU node can be estimated efficiently through the belief propagation algorithm [25].

The DBN inference is similar to the BN inference except for the dynamic transitions. Given the evidences until time  $t: e_1^{1:t}, \dots, e_{17}^{1:t}$ , the posterior probability  $p(AU_1^t, \dots, AU_{17}^t | e_1^{1:t}, \dots, e_{17}^{1:t})$  can be factorized and computed via the AU model by performing the DBN updating process as described in [26].

### 5.4 Convergence of a Knowledge-Driven Model

We employ the sampling scheme described in Section 4.2 to harvest the parameter samples and pseudodata that are consistent with our constraints. To study the convergence of the parameter samples and pseudodata, we have calculated the average SD of the generated parameter instances as a function of the number of parameter samples (as shown in Fig. 8a) and the SD of the model parameters as a function of the size of pseudodata (as shown in Fig. 8b), respectively. For Fig. 8b, we set the quantity of the parameter samples as 1,000 and generate different number of pseudodata from each parameter instance.

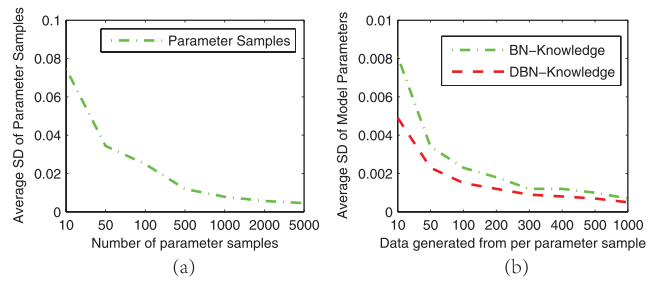


Fig. 8. Convergence of a knowledge-driven model. (a) Convergence of parameter samples. (b) Convergence of data samples.

It is clear from Fig. 8a that the model parameter variation starts stabilizing once the number of parameter samples reaches 1,000. This demonstrates the efficiency of the proposed sampling method. Likewise, Fig. 8b shows that we need generate 500 data from each parameter sample to have a stable estimation of the BN model parameter.

### 5.5 Comparison on Specific Database

We first test the prediction power of the proposed method and compare with that of the data-driven prior model, on specific database, i.e., on C-K database and on FERA database respectively. Similar to the work in [10], a data-driven prior model learns DBN model from the data. Since the prior model should combine with image measurements to infer the true state of each AU, we first extract AU measurements through the AdaBoost classifier. We collect 8,000 images from C-K database and 5,000 images from FERA database, and on both databases, we divide the data into seven sections, each of which contains images from different subjects. We adopt leave-one-fold-out cross validation to evaluate our system.

#### 5.5.1 Comparison with a Data-Driven Prior Model

Given the AU measurements, we fix one section data as testing data and the other six sections as training data for a data-driven prior model. The amount of training data needed to train the data-driven DBN model can be estimated by Hoeffding bound [46]:  $P_D(T_D \notin [p - \varepsilon, p + \varepsilon]) \leq 2e^{-2M\varepsilon^2} < \delta$ , where  $T_D$  is the probability we want to estimate, for example, a parameter in the DBN model,  $p$  is the true probability, and  $M$  is the number of training samples. From Hoeffding bound (setting  $\varepsilon = 0.1$  and  $\delta = 0.01$ ), we can get a minimum  $M = 265$  for one parent configuration. In this work,  $AU_{17}$  node at time  $t$  has eight parent configurations; hence, the amount of minimum training samples needed is  $265 \times 8 = 2,120$ . Since we got more training data than 2,120, we can train a stable data-driven prior model. Fig. 9 shows the comparison results on C-K and FERA database, respectively. We can see from Fig. 9a that when testing on C-K database, DBN learned from knowledge (DBN knowledge) significantly improves the measurements (AdaBoost). The average F1 measure ( $F1 = 2 \frac{P \times R}{P + R}$  where  $P$  is precision and  $R$  is recall) for all AUs increases from 69.76 percent for AdaBoost to 78.09 percent for DBN knowledge. The improvement mainly comes from the AUs that are hard to detect but have strong relationship with other AUs. For instance, the activation of

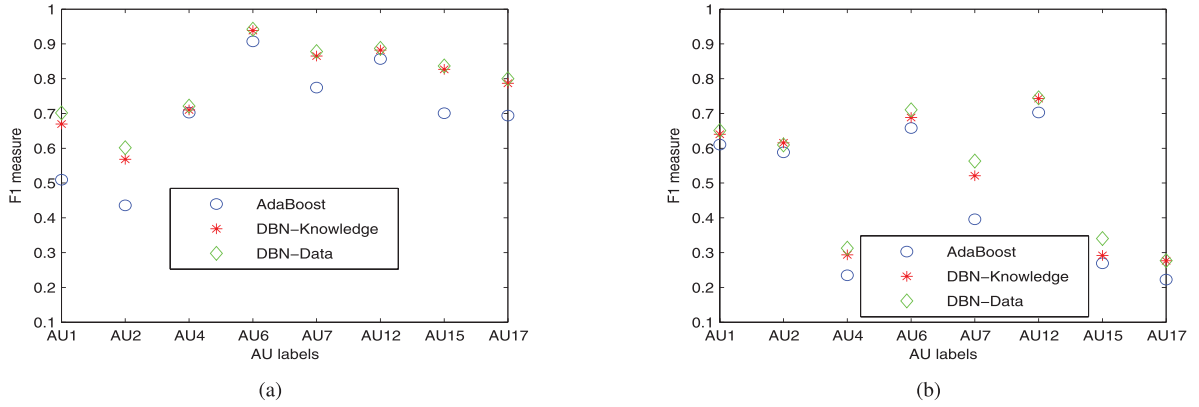


Fig. 9. Comparison of a knowledge-driven prior model with a data-driven prior model on (a) C-K database and (b) FERA database, respectively.

AU1 and AU2 induces less changes in skin texture and are not well recognized by the AdaBoost. Fortunately, the probability of these two action's co-occurrence is high, because they are contracted by the same facial muscle group. By employing such relationship, the DBN knowledge improves the F1 measure of AU1 from 50.96 to 66.95 percent, and that of AU2 from 43.58 to 56.83 percent. Similarly, by employing the co-occurrence relationship between AU15 and AU17, and the coabsence relationship of these two AUs with AU12, the F1 measure of AU15 is increased from 70.09 to 82.68 percent, and that of AU17 is increased from 69.40 to 78.70 percent. Additionally, for comparison, we also evaluate the DBN learned from full training data (DBN data). Its average F1 measure is 79.61 percent, which is slightly better than that of DBN knowledge (78.09). These results are extremely encouraging, as the proposed model uses no training data but domain specific yet generic knowledge to achieve comparable recognition results to DBN learned from full training data. We repeat this experiment on FERA data set as shown in Fig. 9b. The DBN knowledge improves the average recognition results (F1 measure) from 46.03 percent (AdaBoost) to 50.88 percent, and DBN data achieve an average F1 measure of 52.62 percent. Experiments on both data sets prove the prediction power of the proposed method that is practically significant, since in many applications, acquiring the annotated training data is an

expensive, subjective, and time-consuming process, yet there are always plenty of domain knowledge that is often ignored. Note both the data-driven and the knowledge-driven prior model yield improved performance on the C-K database even though its expressions are posed. This is because the constraints we extract on AUs are based mainly on study of facial anatomy and FACS coding. These constraints hence also apply to posed expressions. But the performance improvement should be larger for the non-posed expression since some of the constraints such as the group and dynamic constraints are derived mainly from the spontaneous expression.

### 5.5.2 Comparison with State-of-the-Art Methods

There are lots of works about expression recognition evaluated on C-K database, and Table 5 shows the comparison of the proposed knowledge-driven model with some earlier works. Our results in term of classification rate are better than most previous works. Bartlett et al. [15] and Lucey et al. [31] both achieve high accuracy 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 the AUs are much more difficult to recognize. For a fair comparison, we also evaluate our method only on the initial and the peak frames, and we achieve a classification rate of 97.02 percent, which is better than that in [15] (94.8 percent) and [31]

TABLE 5  
Comparison of Our Work with Some Earlier Works on CK Database

| Author                       | features              | classification  | AUs | CR    | F1    |
|------------------------------|-----------------------|-----------------|-----|-------|-------|
| Bartlett et al. 2005 [15].   | Gabor filters         | AdaBoost+SVM    | 17  | 94.8  |       |
| Bartlett et al. 2006 [21]    | Gabor filters         | AdaBoost+SVM    | 20  | 90.9  |       |
| Whitehill and Omlin 2006 [1] | Haar wavelets         | AdaBoost        | 11  | 92.4  |       |
| Littlewort et al. 2006 [22]  | Gabor filters         | AdaBoost+SVM    | 7   | 92.4  |       |
| Lucey et al. 2007 [31]       | AAM                   | SVM             | 15  | 95.5  |       |
| Valstar & Pantic 2006[34].   | tracked facial points | AdaBoost+SVM    | 15  | 90.2  | 72.9  |
| Tong et al. 2007 [10]        | Gabor filters         | AdaBoost+DBN    | 14  | 93.3  |       |
| Koelstra et al. 2010[32]     | FFD                   | GentleBoost+HMM | 18  | 89.8  | 72.1  |
| Valstar & Pantic 2012 [35]   | tracked facial points | GentleSVM+HMM   | 22  | 91.7  | 59.6  |
| This work                    | Gabor filter          | AdaBoost+DBN    | 8   | 92.78 | 78.09 |
| This work*                   | Gabor filter          | AdaBoost+BN     | 8   | 97.02 | 86.80 |

This work\* means employing BN as prior model and evaluate only on the initial and peak frames  
AU = No. of AUs recognized, CR = Classification Rate, F1 = F1 measure

TABLE 6  
Results for Testing for Eight AUs on CK Data Set

| AUs | F1    | F1[35] | F1[32] | F1[34] |
|-----|-------|--------|--------|--------|
| 1   | 65.95 | 82.6   | 86.89  | 87.6   |
| 2   | 56.83 | 83.3   | 90.00  | 94.0   |
| 4   | 71.01 | 63.0   | 73.13  | 87.4   |
| 6   | 93.83 | 80.0   | 80.00  | 88.0   |
| 7   | 86.52 | 29.0   | 46.75  | 76.9   |
| 12  | 88.24 | 83.6   | 83.72  | 92.1   |
| 15  | 83.68 | 36.1   | 70.27  | 30.0   |
| 17  | 78.70 |        | 76.29  |        |
| Avg | 78.09 | 65.37  | 75.88  | 79.43  |

F1 = F1 measure of our model  
 F1 [35] = F1 Valstar & Pantic 2012[35]  
 F1 [32] = F1 Koelstra et al. 2010[32]  
 F1 [34] = F1 Valstar & Pantic 2006[34]

(95.5 percent). In addition, the classification rate is often less informative, especially when the data are unbalanced. So, we also report our results in term of F1 measure (a harmonic mean of precision and recall rate), which is a more comprehensive metric. From Table 5, we can see that the proposed method significantly outperforms the three earlier works who also reported their results using F1 measure. Since these three works recognize more AUs, we also make a deep comparison on each individual AU as shown in Table 6. On average, our method achieves better or similar results, but it is interesting that these three works get much better results at AU1 and AU2, while our method significantly outperforms them for AU15. Valstar and Pantic [35], [34] employ geometric features, and Koelstra et al. [32] use free-form deformations features that are all powerful to detect AUs such as AU1 and AU2, the activations of which are characterized by large morphological changes but less changes in skin texture. On the other hand, the activation of AU15 involves distinct changes in skin texture without large displacements of facial fiducial points, and hence, Valstar and Pantic [35], [34] fail at AU15. On FERA database, Valstar et al. [33] provided the baseline system for FERA challenge 2011, which employed LBP features and SVM classifier and achieved an average F1 measure of 44.30 percent for the same eight target AUs as in this work, while the proposed knowledge-driven model achieves an average F1 measure of 50.88 percent.

TABLE 7  
Parameters of AU7 Node for Three Different Models

|               | $P(AU7 = 1   AU2 = 0)$ | $P(AU7 = 1   AU2 = 1)$ |
|---------------|------------------------|------------------------|
| DBN-CK        | 0.1705                 | 0.0033                 |
| DBN-FERA      | 0.4267                 | 0.3211                 |
| DBN-Knowledge | 0.3842                 | 0.2475                 |

(Ignoring the Dynamic Dependency)

## 5.6 Comparison across Different Databases

### 5.6.1 Comparison with a Data-Driven Prior Model

In this section, we compare the generalization ability of the proposed knowledge-driven prior model with data-driven prior model on C-K database and on FERA database, respectively. As mentioned above, we have got the AU measurements, and on both databases, we fix one section as testing data. Fig. 10 shows the experimental results.

From Fig. 10a, we can see that when testing on C-K database, DBN knowledge consistently outperforms DBN-FERA (DBN trained on FERA database) on all AUs, and the improvements on some certain AUs are significant. For example, DBN-FERA achieves a F1 measure of 75.54 percent for AU7, and 71.89 percent for AU15, while DBN knowledge achieves a F1 measure of 86.52 percent for AU7, and 82.68 percent for AU15. This means that when prior model trained on FERA applying to C-K data set, it may fail on some certain AUs, i.e., AU7, AU15, vice versa as shown in Fig. 10b. This is because that every data set has its own built-in bias, i.e., the relationship of AU7 with other AUs on FERA data set is not exactly the same as that on C-K data set. This is in particular the case since C-K database consists of posed expression, while FERA data set contains spontaneous expression. At the same time, the DBN knowledge captures the most generic knowledge in the domain, and the parameters of DBN knowledge will not be far from all data sets. To clearly demonstrate this point, we list the parameters of AU7 node (ignoring the dynamic dependency) for three different models in Table 7. From Table 7, we can see that the parameters of DBN knowledge lie between the parameters of DBN-CK and DBN-FERA just as we analyzed.

Though on specific data set, DBN knowledge may achieve a slightly worse result compared to DBN data (DBN model trained on specific data set as shown in

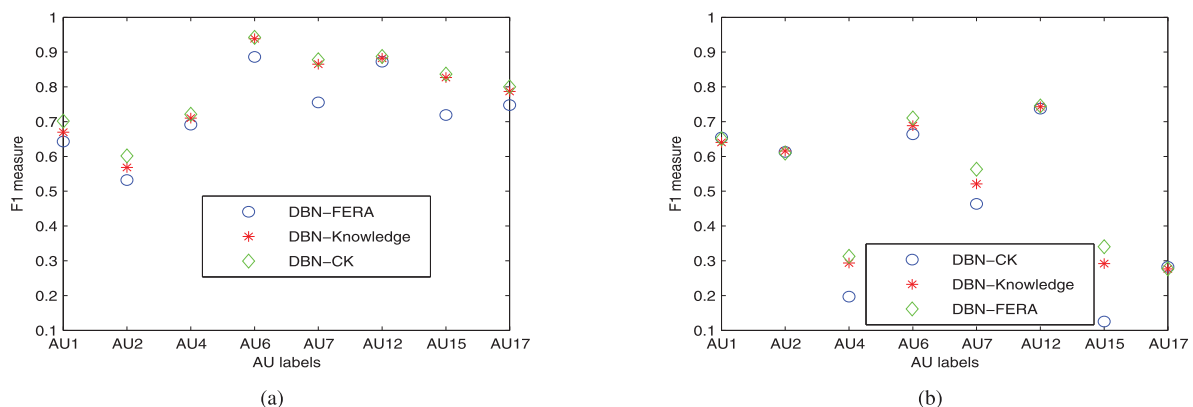


Fig. 10. Comparison of the generalization ability of DBN learned from data and DBN learned from generic knowledge. (a) Test on C-K database. (b) Test on FERA database.

TABLE 8  
Comparison of Knowledge-Driven Model and Data-Driven Model on CK and FERA Data Set, Respectively

| F1 | Test On CK |       |       | Test On FERA |       |       |
|----|------------|-------|-------|--------------|-------|-------|
|    | DBN-F      | DBN-K | DBN-C | DBN-C        | DBN-K | DBN-F |
| F1 | 73.90      | 78.09 | 79.61 | 46.71        | 50.88 | 52.62 |

F1 = F1 measure, DBN-F = DBN-FERA (trained on FERA database)  
DBN-K = DBN-Knowledge, DBN-C = DBN-CK (trained on CK database)

Section 5.5.1), which is also encouraging because we do not use any data for the training purpose of DBN knowledge, when generalizing to different data sets, DBN knowledge significantly outperforms DBN data, which is another benefit of using the knowledge-driven model. Table 8 summarizes the average recognition results of knowledge-driven model and data-driven model on CK and FERA data sets, respectively.

To further compare the knowledge-driven model with data-driven model, we combine the C-K and FERA databases together to train a prior model (DBN combined) and test on C-K and FERA data sets, respectively. Fig. 11 shows the experimental results. From Fig. 11, we can see that combining data from different data sets to train the prior model did not get better results than model trained on the same data set. This is mainly because that each data set has its own built-in bias, and combining data from other data sets will also involve these bias. For instance, DBN combined achieved an average F1 measure of 77.76 percent on CK data set, which is slightly worse than DBN knowledge (78.09 percent) and DBN-CK (79.61 percent). Experiments on FERA data set also show the same fact: DBN combined achieved an average F1 measure of 50.58 percent, while DBN knowledge and DBN-FERA achieved an average F1 measure of 50.88 and 52.62 percent, respectively.

### 5.6.2 Comparison with Image-Driven Methods

In this section, we compare the generalization ability of using image-driven method along and combining image-driven method with the prior model. We first train three kinds image-driven models on C-K database: SVM with liner kernel function (SVM-L-C), SVM with RBF kernel function (SVM-R-C) and AdaBoost (AdaB-C). We use the same feature set: Gabor features selected by AdaBoost.

TABLE 9  
Comparison of Generalization Ability

| (a)          |         |       |         |       |       |       |
|--------------|---------|-------|---------|-------|-------|-------|
| Test On FERA |         |       |         |       |       |       |
| F1           | SVM-L-C | DBN-K | SVM-R-C | DBN-K | AdB-C | DBN-K |
| F1           | 24.56   | 39.28 | 25.51   | 40.62 | 31.15 | 39.01 |

| (b)        |         |       |         |       |       |       |
|------------|---------|-------|---------|-------|-------|-------|
| Test On CK |         |       |         |       |       |       |
| F1         | SVM-L-F | DBN-K | SVM-R-F | DBN-K | AdB-F | DBN-K |
| F1         | 19.92   | 34.27 | 30.95   | 40.28 | 35.38 | 43.00 |

F1 = F1 measure, DBN-K = DBN-Knowledge  
SVM-L-C/F = SVM with liner kernel (trained on CK/FERA database)  
SVM-R-C/F = SVM with RBF kernel (trained on CK/FERA database)  
AdB-C/F = AdaBoost (trained on CK/FERA database)

(a) Train on CK and test on FERA database. (b) Train on FERA and test on CK database.

We test these three models on FERA data set, and since there is large bias between C-K and FERA data sets, all these three image-driven models achieve low recognition results. By combining the low image measurements with the knowledge-driven prior model, we get significant improvements (as shown in Table 9a). Although the final results are still worse than the model trained and tested on the same data set (F1 measure of 46.03 percent), the F1 measure improvement by the prior model is significant. Note that we do not use any FERA data for the training purpose of either the measurement or the prior model. For a complete comparison, we also train image-driven methods on FERA and test on C-K data set and combine the image measurements with the knowledge-driven prior model (as shown in Table 9b). We can reach the same conclusion that combining the prior model can improve the generalization ability of image-drive methods.

## 6 CONCLUSION AND FUTURE WORK

In this work, we propose a knowledge-driven prior model based on a DBN to model the spatial-temporal relationships among AUs to further improve over the image-driven methods, which usually recognize AUs or AU combinations individually and statically. Unlike traditional data-driven prior model, our model is completely learned from generic prior knowledge, which can be expressed as qualitative constraints on individual AUs, on group AUs, and on AU dynamics. We introduce a unified MCMC method to simultaneously incorporate these knowledge into the DBN

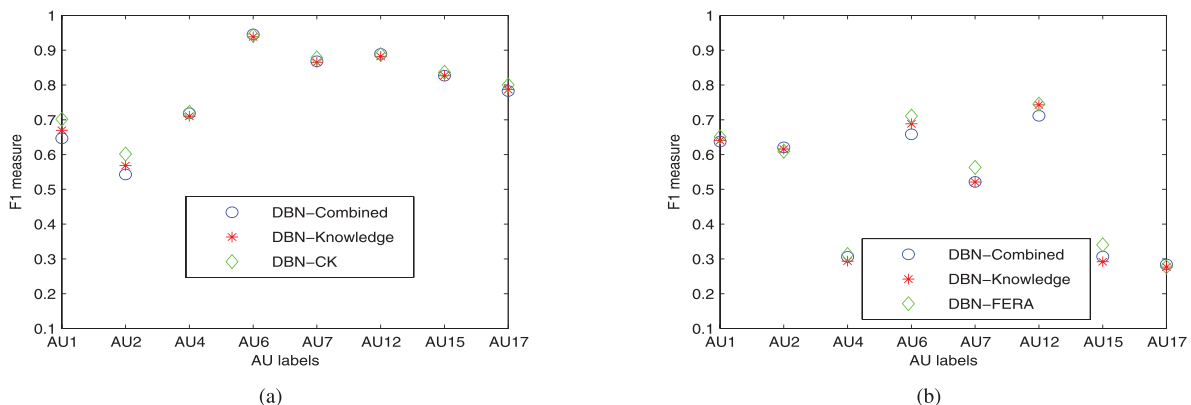


Fig. 11. Comparison of a knowledge-driven model with a data-driven prior model learned from combined data on (a) C-K database and (b) FERA database, respectively.

model learning in a principled manner. As shown in the experiments, the prior model integrated with the feature extraction method yields significant improvement for AU recognition over using a computer vision technique alone. Furthermore, with no training data but generic domain knowledge, the proposed knowledge-driven prior model achieves comparable results to the data-driven prior model for specific database and significantly outperforms the data-driven prior model when generalizing to new data set. While the DBN prior model captures the typical and significant relationships among AUs for a majority of the samples, it may not be consistent with every sample. In fact, it may introduce bias. For those samples inconsistent with the prior model, the prior model may not improve recognition on these samples. While this is a weakness of the proposed prior model, it is in fact the case with any prior models.

In this paper, we have demonstrated the performance of the proposed methods on two databases. In the future, we will further validate their performance on more spontaneous expression databases even though we expect they will work equally well. In addition, we will further study facial anatomy to identify additional knowledge that governs facial muscle movements, with a focus on knowledge that controls the dynamic behavior of facial expressions. Applying this knowledge-driven learning approach to domain adaptation and to other computer vision problems is another future work.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] J. Whitehill and C.W. Omlin, "Haar Features for FACS AU Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 217-222, 2006.
- [2] Y. Chang, C. Hu, and M. Turk, "Probabilistic Expression Analysis on Manifolds," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2004.
- [3] A. Torralba and A.A. Efros, "Unbiased Look at Dataset Bias," *Proc. IEEE Int'l Conf. Computer Vision Pattern and Recognition*, 2011.
- [4] M. Valstar and M. Pantic, "Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics," *Proc. IEEE Int'l Conf. Human-Computer Interaction*, pp. 118-127, 2007.
- [5] R.S. Niculescu, T. Mitchell, and R.B. Rao, "Bayesian Network Learning with Parameter Constraints," *J. Machine Learning Research*, vol. 7, pp. 1357-1383, 2006.
- [6] Y. Mao and G. Lebanon, "Domain Knowledge Uncertainty and Probabilistic Parameter Constraints," *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, 2009.
- [7] Y. Tong and Q. Ji, "Learning Bayesian Networks with Qualitative Constraints," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2008.
- [8] W. Liao and Q. Ji, "Learning Bayesian Network Parameters under Incomplete Data with Qualitative Domain Knowledge," *Pattern Recognition*, vol. 42, pp. 3046-3056, 2009.
- [9] Y. Tong, J. Chen, and Q. Ji, "A Unified Probabilistic Framework for Spontaneous Facial Activity Modeling and Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 258-273, 2010.
- [10] Y. Tong, W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683-1699, Oct. 2007.
- [11] M.H. Mahoor, M. Zhou, K.L. Veon, S. Mavadati, and J. Cohn, "Facial Action Unit Recognition with Sparse Representation," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2011.
- [12] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
- [13] C.P. de Campos and Q. Ji, "Constraints on Priors and Estimations for Learning Bayesian Network Parameters," *Proc. 19th Int'l Conf. Pattern Recognition*, 2008.
- [14] J.J.J. Lien, T. Kanade, J.F. Cohn, and C.C. Li, "Detection, Tracking, and Classification of Action Units in Facial Expression," *Robotics and Autonomous Systems*, vol. 31, pp. 131-146, 2000.
- [15] 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," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.
- [16] Y.L. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, Feb. 2001.
- [17] Y.L. Tian, T. Kanade, and J.F. Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 218-223, 2002.
- [18] P. Ekman and W.V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
- [19] T. Kanade, J.F. Cohn, and Y.L. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 46-53, 2000.
- [20] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Consulting Psychologists Press, 1999.
- [21] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 223-230, 2006.
- [22] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of Facial Expression Extracted Automatically from Video," *Image and Vision Computing*, vol. 24, pp. 615-625, 2006.
- [23] K.L. Schmidt and J.F. Cohn, "Dynamic of Facial Expression: Normative Characteristics and Individual Differences," *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2001.
- [24] T. Kanade, J.F. Cohn, and Y.L. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 46-53, 2000.
- [25] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
- [26] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*. Chapman and Hall/CRC, 2004.
- [27] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge Univ. Press, 1982.
- [28] Social Signal Processing Network, "GEMEP-FERA," <http://sspnet.eu/2011/05/gemep-fera/>, 2013.
- [29] P. Ekman, W.V. Friesen, and J.C. Hager, *Facial Action Coding System: The Manual*. Network Information Research Corp., 2002.
- [30] Y.L. Tian, T. Kanade, and J.F. Cohn, "Facial Expression Analysis," *Handbook of Face Recognition*, S.Z. Li and A.K. Jain, eds., pp. 247-276, Springer, 2005.
- [31] S. Lucey, A. Ashraf, and J.F. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," *Face Recognition*, K. Delac and M. Grgic, eds., pp. 275-286, InTech Education and Publishing, 2007.
- [32] S. Koelstra, M. Pantic, and I. Patras, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940-1954, Nov. 2010.
- [33] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The First Facial Expression Recognition and Analysis Challenge," *Proc. Automatic Face and Gesture Recognition and Workshops*, 2011.
- [34] M. Valstar and M. Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 3, no. 149, 2006.

- [35] M. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 1, pp. 28-43, Feb. 2012.
- [36] M. Valstar, M. Pantic, and I. Patras, "Motion History for Facial Action Detection from Face Video," *Proc. IEEE Conf. Systems, Man, and Cybernetics*, pp. 635-640, 2004.
- [37] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915-928, June 2007.
- [38] Y. Wang, H. Ai, B. Wu, and C. Huang, "Real Time Facial Expression Recognition with AdaBoost," *Proc. 17th Int'l Conf. Pattern Recognition*, 2004.
- [39] C. Shan, S. Gong, and P.W. McOwan, "Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study," *Image and Vision Computing*, vol. 27, pp. 803-816, 2009.
- [40] 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 Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, Jan. 2009.
- [41] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition Both Labeled and Unlabeled Data," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 595-601, 2003.
- [42] S. Gokturk, J. Bouguet, C. Tomasi, and B. Girod, "Model-Based Face Tracking for View Independent Facial Expression Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 272-278, 2002.
- [43] P. Wang and Q. Ji, "Learning Discriminant Features for Multi-View Face and Eye Detection," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 373-379, 2005.
- [44] Nucleus Medical Media, "Muscles of the Face—Medical Illustration, Human Anatomy Drawing," <http://catalog.nucleusinc.com/generateexhibit.php?ID=9300>, 2013.
- [45] D.S. Messinger, W.I. Mattson, M.H. Mahoor, and J.F. Cohn, "The Eyes Have It: Making Positive Expressions More Positive and Negative Expressions More Negative," *Emotion*, vol. 12, pp. 430-436, 2012.
- [46] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *J. Am. Statistical Assoc.*, vol. 58, pp. 13-30, 1963.
- [47] T. Bänziger and K.R. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus," *Blueprint for Affective Computing: A Sourcebook*, K.R. Scherer, T. Bänziger, and E.B. Roesch, eds., chapter 6.1, pp. 271-294, Oxford Univ. Press, 2010.
- [48] Z. Ambadar, J.F. Cohn, and L.I. Reed, "All Smiles Are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous," *J. Nonverbal Behavior*, vol. 33, no. 1, pp. 17-34, 2009.
- [49] D. Keltner and B.N. Buswell, "Embarrassment: Its Distinct Form and Appeasement Functions," *Psychological Bull.*, vol. 122, no. 3, pp. 250-270, 1997.



**Yongqiang Li** received the BS and MS degrees in instrument science and technology from the Harbin Institute of Technology, China, in 2007 and 2009, respectively. He is currently working toward the PhD degree at the Harbin Institute of Technology. He was a visiting student at Rensselaer Polytechnic Institute, Troy, New York, from September 2010 to September 2012. His areas of research include computer vision, pattern recognition, and human-computer interaction.



interaction, and human behavior tracking.

**Jixu Chen** received the BS and MS degrees in electrical engineering from the University of Science and Technology of China in 2003 and 2006, respectively. He received the PhD degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, New York, in 2011. He is currently a researcher at the General Electric Global Research Center, Niskayuna, New York. His areas of research include computer vision, probabilistic graphical model, human-computer



**Yongping Zhao** received the PhD degree in electrical engineering from the Harbin Institute of Technology (HIT), China. He is currently a professor in the Department of Instrument Science and Technology at HIT. His areas of research include signal processing, system integration, and pattern recognition.



**Qiang Ji** received the PhD degree in electrical engineering from the University of Washington. He is currently a professor in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He recently served as a program director at the US National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at the University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, the Department of Computer Science at the University of Nevada, Reno, and the US Air Force Research Laboratory. He currently serves as the director of the Intelligent Systems Laboratory at RPI. His research interests include computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published more than 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including the NSF, the National Institutes of Health, the US Defense Advanced Research Projects Agency, the Office of Naval Research, the Army Research Office, and the Air Force Office of Scientific Research, as well as by major companies including Honda and Boeing. He is an editor of several related IEEE and international journals, and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. He is a fellow of the IAPR.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).