

Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships

Yan Tong, *Student Member, IEEE*, Wenhui Liao, *Member, IEEE*, and Qiang Ji, *Senior Member, IEEE*

Abstract—A system that could automatically analyze the facial actions in real time has applications in a wide range of different fields. However, developing such a system is always challenging due to the richness, ambiguity, and dynamic nature of facial actions. Although a number of research groups attempt to recognize facial action units (AUs) by improving either the facial feature extraction techniques or the AU classification techniques, these methods often recognize AUs or certain AU combinations individually and statically, ignoring the semantic relationships among AUs and the dynamics of AUs. Hence, these approaches cannot always recognize AUs reliably, robustly, and consistently. In this paper, we propose a novel approach that systematically accounts for the relationships among AUs and their temporal evolutions for AU recognition. Specifically, we use a dynamic Bayesian network (DBN) to model the relationships among different AUs. The DBN provides a coherent and unified hierarchical probabilistic framework to represent probabilistic relationships among various AUs and to account for the temporal changes in facial action development. Within our system, robust computer vision techniques are used to obtain AU measurements. Such AU measurements are then applied as evidence to the DBN for inferring various AUs. The experiments show that the integration of AU relationships and AU dynamics with AU measurements yields significant improvement of AU recognition, especially for spontaneous facial expressions and under more realistic environment including illumination variation, face pose variation, and occlusion.

Index Terms—Facial action unit recognition, facial expression analysis, Facial Action Coding System, Bayesian networks.

1 INTRODUCTION

FACIAL expressions are the most powerful and natural means of communication among human beings. A large percentage of the approaches on facial expression analysis attempt to recognize six basic facial expressions such as happiness, sadness, surprise, disgust, fear, and anger. However, these basic expressions only represent a small set of human facial expressions. In fact, human emotion is composed of thousands of expressions, although most of them differ in subtle changes of a few facial features. The Facial Action Coding System (FACS) developed by Ekman and Friesen [1] is the most commonly used system for facial behavior analysis. Based on FACS, the facial behavior is decomposed into 46 action units (AUs), each of which is anatomically related to the contraction of a specific set of facial muscles. Although they only define a small number of distinctive AUs, over 7,000 different AU combinations have been observed so far [2]. 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. However, the effort for training human experts and manually scoring the AUs is expensive and time consuming, and the reliability of manually coding AUs is inherently attenuated by the subjectivity of human coder. Therefore, a system that can recognize AUs in real time without human intervention is more desirable for various application fields including automated tools for behavioral

research, videoconference, affective computing, perceptual human-machine interfaces, 3D face reconstruction and animation, and others.

In general, an automatic AU recognition system consists of two key stages: the facial feature extraction stage and the AU classification stage. A great amount of methods have been devoted to these two stages. In the facial feature extraction stage, the previous work can be summarized into two major groups: holistic techniques [3], [4], [5], [6], [7], [8], where the face is processed as a whole, and local techniques [3], [4], [9], [10], [11], [12], [13], [14], [15], [16], in which only a set of specific facial features or facial regions are considered. The approaches on AU classification could be grouped into spatial approaches [17], [3], [4], [10], [7], [5], [6], [11], [12], [16], [8], [13], [18], where AUs are analyzed frame by frame, and spatial-temporal approaches [3], [9], [14], [15], [19], where AUs are recognized over time based on the temporal evolution of facial features. Basically, these methods differ in either the feature extraction techniques or the AU classification techniques, or both. However, they all classify each AU or certain AU combinations independently and statically (although some of them analyze the temporal properties of facial features), ignoring the semantic relationships among AUs and the dynamics of AUs. Hence, these approaches cannot always recognize AUs reliably, robustly, and consistently due to the richness, ambiguity, and dynamic nature of facial actions, as well as due to the uncertainties with feature extraction. This motivates us to exploit the relationships among the AUs, as well as the temporal development of each AU, in order to improve AU recognition.

Specifically, it is rare that a single AU occurs alone in spontaneous facial behavior. Instead, some patterns of AU combinations appear frequently to express natural human emotions. Thus, some of the AUs would appear simultaneously most of the time. On the other hand, it is

• The authors are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590. E-mail: {tongy2, liaow, jiq}@rpi.edu.

Manuscript received 4 Apr. 2006; revised 10 Oct. 2006; accepted 19 Dec. 2006; published online 18 Jan. 2007.

Recommended for acceptance by T. Darrell.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0258-0406. Digital Object Identifier no. 10.1109/TPAMI.2007.1094.

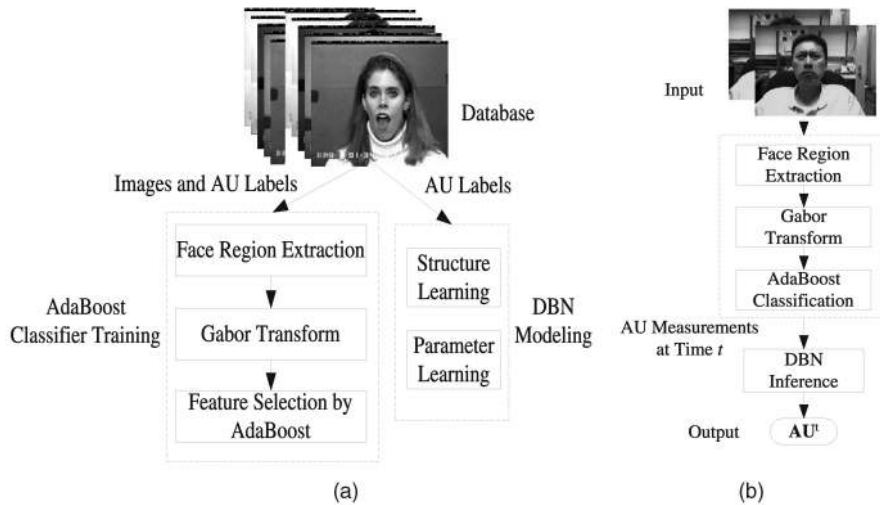


Fig. 1. The flowchart of the real-time AU recognition system. (a) The offline system training for AdaBoost classifiers and DBN. (b) The online AU recognition.

nearly impossible to perform some AUs simultaneously, as described in the alternative rules defined in FACS [20]. Furthermore, in addition to the dynamic development of each AU, the relationships among AUs undergo a temporal evolution to realistically reflect the evolution from a neutral state to a weak emotion, then to the apex, and, finally, to a releasing state. Such dynamic and semantic co-occurrence/coabsence relationships can be well modeled with a dynamic Bayesian network (DBN). Such a model is capable of representing the relationships among different AUs in a coherent and unified hierarchical structure, accounting for uncertainties in the AU recognition process with conditional dependence links, modeling the dynamics in facial behavior development with temporal links, and providing principled inference and learning techniques to systematically combine domain information and statistical information extracted from the data.

Fig. 1 gives the flowchart of our automatic AU recognition system that consists of an offline training phase (Fig. 1a) and an online AU recognition phase (Fig. 1b). The system training includes training an AdaBoost classifier for each AU and learning the DBN in order to correctly model the AU relationships. Advanced learning techniques are applied to learn both the structure and parameters of the DBN based on both training data and domain knowledge. The online AU recognition consists of two independent but collaborative components: AU measurement extraction by AdaBoost classification and DBN inference. First, the face and eyes are detected in live video automatically. Then, the face region is divided into upper and lower parts, which are aligned based on the detected eye positions and are convolved with a set of multiscale and multiorientation Gabor filters, respectively, to produce wavelet features for each pixel. After that, the AdaBoost classifier combines the wavelet features to produce a measurement score for each AU. These scores are then used as evidence by a DBN for AU inference. Under such a framework, the erroneous AU measurements from the underlying computer vision techniques could be compensated by the relationships among AUs embedded in the DBN, and some AUs that are difficult to be directly recognized can be inferred indirectly from other AUs. The experiments show that AU recognition with the proposed DBN is improved

significantly for all target AUs, especially in spontaneous facial expressions and under more realistic conditions such as illumination variation, face pose variation, and occlusion.

2 RELATED WORK

A number of approaches in computer vision have been developed to recognize AUs and their combinations from video sequences or from static images by observing the representative facial appearance changes, as described in FACS. Generally, a two-stage procedure is utilized: First, the features related to the corresponding AUs are extracted, and then, the AUs are recognized by different recognition engines based on the extracted features. In this section, we will present a brief review on the previous approaches based on the two processing stages.

2.1 Facial Feature Extraction

In general, the facial feature extraction techniques could be grouped into two classes: holistic approaches, where the face is processed as a whole, and local approaches, in which only a set of specific facial features or facial regions are taken into account according to the target AU.

2.1.1 Holistic Approaches

Among the holistic facial feature extraction methods, there are three major variations: motion extraction by dense flow estimation [3], [4], motion extraction based on difference images [4], [5], [6], [7], and single-image-based methods [21], [8].

Facial muscular activities would produce the changes in the direction and magnitude of the skin surface motion. Dense optical flow are used in [3], [4] to detect the direction and magnitude of the facial motions related to some AUs by extracting a motion field for the whole face image in each frame. However, extracting dense flow is susceptible to inaccurate image alignment, motion discontinuities, and illumination variation. Furthermore, the computation of dense flow is slow, which prevents it from real-time applications.

Facial motion could also be extracted through the difference between the reference neutral face image and

the testing face image. Further operations are performed on the difference image to extract the image features including the Gabor-wavelet-based feature [6], [7] and low-dimensional principal component coefficients [4], [5], [7]. The difference-image-based methods have the advantage of being robust to the illumination and skin color variations but cannot extract the motion direction.

Besides extracting the holistic facial motion in the image sequence, single-image-based approaches are employed to recognize AUs from the still image. Lanitis et al. [21] utilize statistical learning of images by an Active Appearance model and classify the six expressions based on the appearance parameters generated by the principal components analysis (PCA). Bartlett et al. [8] convolve the whole face images by a set of Gabor wavelet kernels, and the resulting Gabor wavelet magnitudes are used as the input to a recognition engine. This category of feature extraction methods has the advantages of detecting the changes both in geometrical positions of features and in face appearance (that is, wrinkles and bulges) simultaneously and of its generalization ability to recognize a full range of facial behavior. However, the image-based methods have the limitation that only static cues from still face images are explored, without considering the temporal evolution of facial actions.

2.1.2 Local Approaches

Facial muscular activities would also produce the changes in the appearance and positions of permanent facial features and in the presence of transient facial features. Especially, the motion information extracted from the eyebrows, eyes, and mouth is important for recognizing many AUs [22]. The local methods [4], [9], [10], [3], [11], [12], [13], [14], [15], [16] focus on the facial features locally. The facial motion and facial deformation are extracted by tracking a small set of predefined feature points [11], [12], [13], [10], [9], [23], [18] that are located usually around the permanent facial components. To measure the displacement between the current face image and the referenced neutral face consistently, the positions of all tracked facial feature points in each frame should be normalized by mapping them into a standard face. Moreover, the transient facial features are extracted [11], [12], [18] by detecting the high-gradient components for furrows and wrinkles.

The extracted facial features are then quantized into a set of symbolic representations (for example, the inner brow moves up, and the lip height increases) or parametric representations (for example, the size of the facial component, the feature motions, and the existence of the transient features), which is then used for AU recognition.

Rather than extracting facial motion, a small set of shape parameters is generated from the tracked facial feature points by PCA and is utilized in [16], [24] to represent the variations of facial deformations according to facial expressions for AU analysis. However, these methods are not sufficient for identifying the full range of the AUs. For example, AU6 (cheek raiser) cannot be described by facial deformations [24], [13]. Moreover, they have inherent shortcomings including the invalidity of the linear assumption of PCA and the difficulty to distinguish the variation due to facial actions from those caused by the change in illumination, face pose, face size, and individual differences.

In summary, the holistic approaches include more information from a large face region and, thus, have the

generalization ability to recognize a full range of facial behaviors; however, its computational complexity is considerable. The local approaches are sensitive to subtle changes in small areas and more computationally efficient than the holistic approaches, but specially purposed features should be deliberately designed for each AU, respectively. Since the accuracy of the AU recognition fully depends on how reliably and accurately the facial features are extracted, human intervention is usually required in the initial frame to ensure a reliable reference for the subsequent frames in most of the existing facial feature extraction techniques.

2.2 AU Classification

Given the extracted facial features, the AUs are identified by recognition engines. In general, they can be divided into two groups: spatial classification approaches and spatial-temporal classification approaches.

2.2.1 Spatial Classification Approaches

The simplest classification method is identifying AUs based on some parametric rules of AUs [17] or by template matching [4]. However, it is not effective to recognize an AU in the presence of other AUs, since the template would be very different for the AU combinations. Cohn et al. [10] and Lien et al. [3] employ discriminant analysis for AU recognition. Given the measurement of the facial features, the discrimination between AUs is performed by comparing posterior probabilities of AUs.

Neural network is widely used in AU classification [4], [5], [6], [7], [11], [12]. They differ mainly in the input facial features, which could be PCA coefficients [4], [5], [7], facial feature motions and transient facial features [4], [11], and Gabor wavelet coefficients [6], [12]. Multiple AU outputs could be enabled for detecting AUs simultaneously in an AU combination so that a target AU could be identified in the presence of other AUs. However, training the neural network is difficult for unconstrained (spontaneous) facial behavior, in which over thousands of AU combinations have been observed.

Support Vector Machines (SVMs) have been used for identifying AUs in [16], [8], [13]. Bartlett et al. [8] investigate machine learning techniques including SVMs, Linear Discriminant Analysis, and AdaBoost. The best recognition performance is obtained through SVM classification on a set of Gabor wavelet coefficients selected by AdaBoost. However, the computational complexity of SVMs is considerable.

One limitation of the spatial classification methods is that they do not model the temporal dynamics of a facial behavior, although the temporal evolution often reveals more information about the underlying human emotional states [14]. Another common limitation of such methods is that they only attempt to classify each AU independently, ignoring the semantic relationships among AUs. Recently, some researchers begin to realize the problem and try to use a few simple relationships to help AU recognition. Pantic and Rothkrantz [18] utilize a rule-based method with fast direct chaining inference procedure for AU recognition. Based on the description of each AU in the FACS system, a set of rules are generated in terms of the feature representation, as well as a few simple relationships among AUs. For example, the rule for classifying AU6 is, "If AU12 (lip corner raiser) or AU13 (sharp lip puller), then AU6." However, some of the rules are not correct for more general cases. For example, although a

strong AU12 could cause the contraction of AU6, a weak action of AU12 could be contracted without adding AU6. Moreover, it is not clear how these rules handle the uncertainty associated with AU measurements. Therefore, probabilistically modeling the relationships among AUs would be more desirable than enumerating some simple deterministic rules.

2.2.2 Spatial-Temporal Classification Approaches

There are several attempts to recognize AUs over time. Lien et al. [3] recognize AUs based on a set of Hidden Markov Models (HMMs). Each AU or AU combination is assigned an HMM. The classification is performed by choosing the AU or AU combination, which maximizes the likelihood of the extracted facial features generated by the associated HMM. However, the large number of HMMs required to identify a great number of potential AU combinations prevents it from real-time applications.

In addition to the above approaches, there are a few groups using Bayesian networks (BNs) for facial expression classification. The research group of Cohen et al. [9] uses Gaussian Tree-Augmented Naive Bayes (TAN) to learn the dependencies among different facial motion features in order to classify facial expressions. However, due to TAN's structure limitations, it cannot handle complex relationships between facial features, as well as temporal changes. Zhang and Ji [14] exploit a BN to classify six basic facial expressions with a dynamic fusion strategy. Gu and Ji [15] use a similar idea for facial event classification such as fatigue. Cohen et al. [19] further propose a classification-driven stochastic structure search (SSS) algorithm for learning a BN classifier to recognize facial expression from both labeled and unlabeled data.

However, in the previous BN-based facial expression recognition, AUs are modeled as hidden nodes connecting facial features and facial expressions and are not recognized explicitly in their models. Again, they do not consider the semantic relationships among AUs, as well as the dynamics of each AU.

Overall, the previous work on facial AU analysis, in our view, have the following limitations:

- The previous work based on static images could not model the temporal dynamics and the momentary intensity of an observed facial behavior. However, the psychological experiment [25] suggests that the facial behavior is more accurately recognized from an image sequence than from a still image.
- Existing work based on spatial-temporal analysis requires starting from a neutral expression. Since facial motion information critically depends on the accuracy of image registration between the reference (neutral) face and the target face, human intervention in the initial frame is inevitable to ensure a reliable reference image for the subsequent frames. A more automated system is desired.
- Most of the previous work recognizes each AU combination as a new AU. However, it is nearly impossible to deal with the potential thousands of AU combinations.
- The missing features caused by occlusions, measurement errors, and low AU intensity have not been addressed effectively by the existing work.

- The existing methods only attempt to classify each AU independently or statically, ignoring the semantic relationships among AUs.

Our system differs from the cited ones in that it explicitly models probabilistic relationships among different AUs and accounts for the temporal changes in facial action development with a DBN. Advanced learning techniques are used to learn both the structure and the parameters of the DBN from the training database. To the best of our knowledge, we are the first to exploit the relationships among AUs in such a systematic and unified model to improve AU recognition, in addition to advanced computer vision techniques.

3 AU MEASUREMENT EXTRACTION

3.1 Face and Eyes Detection















As described above, the accuracy of face alignment would affect the performance of AU recognition. Since the position, size, and orientation of the face region are usually estimated using the knowledge of eye centers, an accurate and robust eye detector is desirable. In this work, a boosted eye detection algorithm is employed based on the recursive nonparametric discriminant analysis (RNDA) features proposed in [26]. For eye detection, the RNDA features provide better accuracy than Harr features [27] especially for nonfrontal faces since they are not constrained with rectangular shape. The features are sequentially learned and combined with AdaBoost to form an eye detector. To improve the speed, a cascade structure is applied. The eye detector is trained on thousands of eye images and more noneye images. The resulting eye detector uses less than 100 features.

The eye localization follows a hierarchical principle: First, a face is detected, and then, the eyes are located inside the detected face. The detection rate for frontal view face is around 95.0 percent, and the eye detection rate on the detected face is about 99.0 percent. An overall 94.5 percent eye detection rate is achieved, with 2.67 percent average normalized error (the pixel error normalized by the distance between two eyes) [28] on Face Recognition Grand Challenge (FRGC) 1.0 database [29]. Given the knowledge of eye centers, the face region is divided into upper and lower face parts, each of which is normalized and scaled into a 64×64 image.

3.2 AU Measurements

For subsequent discussions, Table 1 summarizes a list of commonly occurring AUs and their interpretations although the proposed system is not restricted to recognizing these AUs, given the training data set. Since the upper/lower facial actions are contracted, respectively, by the related upper/lower facial muscles [1], we extract the measurements for the upper/lower AUs on the upper/lower face region separately. Given the normalized upper/lower face image, we extract a measurement for each AU through a general-purpose learning mechanism based on Gabor feature representation and AdaBoost classification. For this work, the magnitudes of a set of multiscale and multiorientation Gabor wavelets as in [30] are used as the feature representation, similar to the work by Bartlett et al. [8]. Gabor-wavelet-based feature representation has the psychophysical basis of human vision and achieves robust performance for expression recognition and feature extraction under illumination and appearance variations [31], [12]. Instead of extracting the

TABLE 1
A List of AUs and Their Interpretations (Adapted from [14])

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

Gabor wavelet features at specific feature points, the whole normalized upper/lower face region is convolved pixel by pixel by a set of Gabor filters at five spatial frequencies and eight orientations. Each Gabor wavelet coefficient J is represented by a complex number $J = \|J\|e^{j\phi}$ with magnitude $\|J\|$ and phase ϕ . The phase changes drastically to small location displacement. Hence, it is susceptible to inaccurate image alignment if the phase is used for feature representation. On the other hand, the magnitude changes slowly as location changes. Therefore, only the magnitudes of Gabor wavelet coefficients are employed for feature representations. Overall, $8 \times 5 \times 64 \times 64 = 163,840$ individual Gabor wavelet features are produced as the input of each AdaBoost classifier. This feature extraction method has the advantage that it captures not only the changes in geometrical positions of permanent facial components, but also the changes in facial appearance (that is, wrinkles and bulges) encoded by the Gabor wavelet coefficients. Furthermore, this feature extraction method could be generalized to recognize a full range of facial behavior.

The AdaBoost classifier is then employed to obtain the measurement for each AU. AdaBoost is not only a good feature selection method, but also a fast classifier. In this work, feature selection and classifier construction are performed simultaneously by AdaBoost. At first, each training sample is initialized with an equal weight. Then, weak classifiers are constructed using the individual Gabor features during AdaBoost training. In each iteration, the single weak classifier (Gabor wavelet feature) with the minimum weighted AU classification error with respect to the current boosting distribution is selected. Then, it is linearly combined with the previously selected weak classifiers to form the final “strong” classifier with a weight that is proportional to the minimum error. The training samples are then reweighted based on the performance of the selected weak classifier, and the process is repeated. Since the weights of the wrongly classified examples are increased in each iteration, AdaBoost forces the classifier to focus on the most difficult samples in the training set, and thus, it results in an effective classifier. To increase the speed of the actual algorithm, the final classifier is broken up into a series of cascaded AdaBoost classifiers. A large majority of the negative samples will be removed in earlier cascades. Therefore, it results in a much faster real-time classifier. In our work, the final classifier utilizes around 200 Gabor features for each AU. Compared with the AU

classification by SVMs, the AdaBoost classifier significantly improves the recognition efficiency [8].

However, this image-appearance-based approach treats each AU and each frame individually and heavily relies on the accuracy of face region alignment. In order to model the dynamics of AUs, as well as their semantic relationships, and to deal with the image uncertainty, we utilize a DBN for AU inference. Consequently, the continuous output of the AdaBoost classifier is discretized into binary values and used as the evidence for the subsequent AU inference via the DBN.

4 AU INFERENCE WITH A DBN

4.1 AU Relationship Analysis

As discussed before, due to the richness, ambiguity, and dynamic nature of facial actions, as well as the image uncertainty and individual difference, current computer vision methods cannot perform feature extraction reliably, which limits AU recognition accuracy. Moreover, when AUs occur in a combination, they may be nonadditive; that is, the appearance of an AU in a combination is different from its appearance of occurring alone. Fig. 2 demonstrates an example of the nonadditive effect: when AU12 appears alone, the lip corners are pulled up toward the cheekbone; however, if AU15 (lip corner depressor) is also present, then the lip corners are somewhat angled down due to the action of AU15. The nonadditive effect increases the difficulty of recognizing AUs individually. Fortunately, there are some inherent relationships among AUs, as described in the FACS manual [20]. For example, the alternative rules provided in the FACS manual describe the mutual exclusive relationship among



Fig. 2. Nonadditive effect in an AU combination (reprinted with permission from Ekman et al. [20]). (a) AU12 occurs alone. (b) AU15 occurs alone. (c) AU12 and AU15 appear together.



Fig. 3. Some examples of AU combinations show meaningful expressions (adapted from [32]). (a) AU6+AU12+AU25 represents the happiness facial expression. (b) AU1+AU2+AU5+AU25+AU27 represents the surprise facial expression. (c) AU1+AU4+AU15+AU17 represents the sadness facial expression.

AUs. The two AUs are “alternative,” since “it may not be possible anatomically to do both AUs simultaneously” or “the logic of FACS precludes the scoring of both AUs” [20]. Furthermore, FACS also includes the co-occurrence rules in their old version [1], which were “designed to make scoring more deterministic and conservative, and more reliable” [20]. Therefore, if we could provide a unified framework, which could model and learn the relationships among AUs, and employ the relationships in AU recognition automatically, then it will compensate the disadvantages of computer vision techniques and, thus, improve the recognition accuracy. Continuing with the nonadditive example, since it is most likely that AU6 and/or AU17 (chin raiser) are also present together with AU12+AU15, although individually identifying AU12 and AU15 is difficult, they (AU12 and AU15) could be recognized more accurately through the relationships with AU6 and AU17.

In a spontaneous facial behavior, it is rare that a single AU appears alone. Groups of AUs usually appear together to show meaningful facial expressions. For example, the happiness expression may involve AU6+AU12+AU25 (lips part), as in Fig. 3a; the surprise expression may involve AU1 (inner brow raiser)+AU2 (outer brow raiser)+AU5 (upper lid raiser) + AU25 + AU27 (mouth stretch), as in Fig. 3b; and the sadness expression may involve AU1+AU4 (brow lowerer)+AU15+AU17, as in Fig. 3c [14]. Furthermore, some of the AUs would appear simultaneously most of the time. For example, it is difficult to do AU2 without performing AU1, since the activation of AU2 tends to raise the inner portion of the eyebrow as well [18]. AU27 rarely appears without AU25. On the other hand, it is nearly impossible for some AUs to appear together, as in the alternative rules described in the FACS manual [20]. For example, the lips cannot be parted as AU25 and pressed as AU24 (lip presser) simultaneously. In addition to the dynamic development of each AU, the relationships among AUs also undergo a temporal evolution, which can realistically reflect the evolution from a neutral state to a weak emotion, then to the apex and, finally, to a releasing state. For example, for a smile, usually, AU12 is first contracted to express a slight emotion, then with the increasing of emotion intensity, AU6 and/or AU25 would be contracted, and after the actions reach their peak simultaneously, AU6 and/or AU25 would gradually be relaxed and, finally, AU12 would be released, and all AUs return to the neutral state.

4.2 AU Relationships Modeling

Due to the errors from feature extraction and the variability among individuals, the relationships among AUs and their

dynamics are stochastic. We propose to use a BN to model and learn such relationships described above. A BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables. In a BN, nodes denote variables and the links among nodes denote the conditional dependency among variables. The dependency is characterized by a conditional probability table (CPT) for each node.

We derive an initial BN structure by analyzing the AU relationships in FACS-coded images from two facial expression databases: the Cohn-Kanade *DFAT-504* database [32] and the Intelligent Systems Lab (ISL) database constructed by our own research group. We will discuss the two databases in detail in the experimental validation.

The AU relationships are partly learned from the co-occurrence table in Table 2a and the coabsence table in Table 2b. Both tables are constructed from the two facial expression databases. For Table 2a, each entry $a_{i,j}$ represents the probability $P(AU_i = 1|AU_j = 1)$, and the pairwise co-occurrence dependency between two AUs is computed as follows:

$$P(AU_i = 1|AU_j = 1) = \frac{N_{AU_i+AU_j}}{N_{AU_j}}, \quad (1)$$

where $N_{AU_i+AU_j}$ is the total number of the positive examples of the AU combination ($AU_i + AU_j$) regardless of the presence of other AUs in the databases, and N_{AU_j} is the total number of the positive examples of AU_j in the databases.

Similarly, for Table 2b, each entry $a_{i,j}$ represents the probability $P(AU_i = 0|AU_j = 0)$, and the pairwise coabsence dependency between two AUs is computed as follows:

$$P(AU_i = 0|AU_j = 0) = \frac{M_{\neg AU_i+\neg AU_j}}{M_{\neg AU_j}} \quad (2)$$

where $M_{\neg AU_i+\neg AU_j}$ is the total number of the events that neither AU_i nor AU_j occurs, and $M_{\neg AU_j}$ is the total number of the negative examples of AU_j in the databases.

Although some AUs in the upper face are relatively independent from the AUs in the lower face, it is not necessary to learn the tables for the upper and lower face AUs separately, since the moderate values of $P(AU_i = 1|AU_j = 1)$ and $P(AU_i = 0|AU_j = 0)$ imply the relative independency between the AUs. For example, $P(AU1 = 1|AU2 = 1) = 0.988$ in Table 2a means that if AU2 is activated, then it is most likely that AU1 is also contracted. On the contrary, $P(AU9 = 1|AU5 = 1) = 0.002$ means that AU5 and AU9 (nose wrinkler) are almost impossible to occur together. Thus, if the probability is greater than T_{up} or less than T_{bottom} (T_{up} and T_{bottom} are the predefined thresholds), then we assume that the two AUs are strongly dependent, which could be modeled with a link between them in the BN.

Furthermore, in Table 2, we can also see that the dependency between two AUs is not symmetric, that is, $P(AU_i = 1|AU_j = 1) \neq P(AU_j = 1|AU_i = 1)$. This asymmetry indicates that one AU plays the dominant role (the cause), whereas the other AU follows (that is, the effect). This kind of relationship justifies the use of a directed edge instead of an undirected edge to represent the causal dependency between two AUs. This way, an initial BN structure is manually constructed, as in Fig. 4a, to model the relationships among the 14 target AUs.

TABLE 2
(a) The Co-Occurrence Table (Each Entry $a_{i,j}$ Represents $P(AU_i = 1|AU_j = 1)$) and
(b) the Coabsence Table (Each Entry $a_{i,j}$ Represents $P(AU_i = 0|AU_j = 0)$)

| | AU1=1 | AU2=1 | AU4=1 | AU5=1 | AU6=1 | AU7=1 | AU9=1 | AU12=1 | AU15=1 | AU17=1 | AU23=1 | AU24=1 | AU25=1 | AU27=1 |
|--------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|--------|
| AU1=1 | 1.000 | 0.988 | 0.190 | 0.907 | 0.047 | 0.072 | 0.028 | 0.118 | 0.187 | 0.155 | 0.052 | 0.058 | 0.352 | 0.777 |
| AU2=1 | 0.831 | 1.000 | 0.076 | 0.888 | 0.024 | 0.028 | 0.028 | 0.101 | 0.081 | 0.072 | 0.022 | 0.026 | 0.304 | 0.777 |
| AU4=1 | 0.239 | 0.115 | 1.000 | 0.098 | 0.328 | 0.834 | 0.910 | 0.123 | 0.507 | 0.591 | 0.504 | 0.497 | 0.151 | 0.021 |
| AU5=1 | 0.610 | 0.719 | 0.052 | 1.000 | 0.019 | 0.025 | 0.003 | 0.085 | 0.048 | 0.043 | 0.022 | 0.017 | 0.254 | 0.648 |
| AU6=1 | 0.032 | 0.019 | 0.176 | 0.019 | 1.000 | 0.301 | 0.303 | 0.570 | 0.117 | 0.131 | 0.118 | 0.163 | 0.269 | 0.001 |
| AU7=1 | 0.074 | 0.034 | 0.671 | 0.037 | 0.453 | 1.000 | 0.928 | 0.180 | 0.341 | 0.426 | 0.410 | 0.431 | 0.136 | 0.002 |
| AU9=1 | 0.010 | 0.013 | 0.267 | 0.002 | 0.167 | 0.338 | 1.000 | 0.017 | 0.228 | 0.263 | 0.077 | 0.054 | 0.032 | 0.001 |
| AU12=1 | 0.094 | 0.097 | 0.078 | 0.100 | 0.674 | 0.142 | 0.038 | 1.000 | 0.029 | 0.039 | 0.129 | 0.168 | 0.295 | 0.076 |
| AU15=1 | 0.123 | 0.064 | 0.264 | 0.046 | 0.114 | 0.220 | 0.404 | 0.024 | 1.000 | 0.661 | 0.424 | 0.479 | 0.005 | 0.007 |
| AU17=1 | 0.152 | 0.085 | 0.460 | 0.062 | 0.190 | 0.412 | 0.697 | 0.048 | 0.990 | 1.000 | 0.812 | 0.826 | 0.018 | 0.001 |
| AU23=1 | 0.020 | 0.010 | 0.154 | 0.013 | 0.067 | 0.156 | 0.080 | 0.063 | 0.249 | 0.319 | 1.000 | 0.766 | 0.009 | 0.002 |
| AU24=1 | 0.027 | 0.014 | 0.180 | 0.011 | 0.111 | 0.194 | 0.066 | 0.097 | 0.334 | 0.385 | 0.910 | 1.000 | 0.006 | 0.001 |
| AU25=1 | 0.533 | 0.554 | 0.181 | 0.570 | 0.602 | 0.202 | 0.131 | 0.560 | 0.013 | 0.027 | 0.034 | 0.021 | 1.000 | 0.998 |
| AU27=1 | 0.298 | 0.358 | 0.006 | 0.369 | 0.001 | 0.001 | 0.001 | 0.036 | 0.004 | 0.001 | 0.002 | 0.001 | 0.253 | 1.000 |

(a)

| | AU1=0 | AU2=0 | AU4=0 | AU5=0 | AU6=0 | AU7=0 | AU9=0 | AU12=0 | AU15=0 | AU17=0 | AU23=0 | AU24=0 | AU25=0 | AU27=0 |
|--------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|--------|
| AU1=0 | 1.000 | 0.952 | 0.756 | 0.895 | 0.738 | 0.724 | 0.753 | 0.747 | 0.764 | 0.750 | 0.755 | 0.752 | 0.837 | 0.824 |
| AU2=0 | 0.992 | 1.000 | 0.764 | 0.937 | 0.780 | 0.761 | 0.795 | 0.790 | 0.791 | 0.776 | 0.794 | 0.791 | 0.871 | 0.866 |
| AU4=0 | 0.697 | 0.671 | 1.000 | 0.677 | 0.719 | 0.877 | 0.769 | 0.675 | 0.750 | 0.799 | 0.733 | 0.736 | 0.639 | 0.686 |
| AU5=0 | 0.981 | 0.979 | 0.805 | 1.000 | 0.822 | 0.807 | 0.832 | 0.831 | 0.828 | 0.814 | 0.834 | 0.830 | 0.899 | 0.894 |
| AU6=0 | 0.810 | 0.814 | 0.855 | 0.821 | 1.000 | 0.890 | 0.860 | 0.939 | 0.839 | 0.839 | 0.842 | 0.847 | 0.906 | 0.831 |
| AU7=0 | 0.721 | 0.721 | 0.946 | 0.733 | 0.808 | 1.000 | 0.832 | 0.756 | 0.787 | 0.824 | 0.785 | 0.791 | 0.717 | 0.746 |
| AU9=0 | 0.893 | 0.898 | 0.989 | 0.900 | 0.930 | 0.992 | 1.000 | 0.900 | 0.941 | 0.967 | 0.915 | 0.912 | 0.888 | 0.907 |
| AU12=0 | 0.792 | 0.798 | 0.776 | 0.803 | 0.907 | 0.806 | 0.804 | 1.000 | 0.791 | 0.776 | 0.813 | 0.816 | 0.878 | 0.808 |
| AU15=0 | 0.842 | 0.830 | 0.896 | 0.831 | 0.843 | 0.871 | 0.874 | 0.822 | 1.000 | 0.998 | 0.877 | 0.888 | 0.774 | 0.836 |
| AU17=0 | 0.754 | 0.743 | 0.871 | 0.746 | 0.769 | 0.832 | 0.819 | 0.736 | 0.910 | 1.000 | 0.832 | 0.846 | 0.666 | 0.754 |
| AU23=0 | 0.892 | 0.894 | 0.939 | 0.898 | 0.908 | 0.932 | 0.911 | 0.906 | 0.940 | 0.979 | 1.000 | 0.991 | 0.870 | 0.904 |
| AU24=0 | 0.872 | 0.874 | 0.926 | 0.878 | 0.896 | 0.922 | 0.892 | 0.894 | 0.936 | 0.977 | 0.973 | 1.000 | 0.843 | 0.885 |
| AU25=0 | 0.710 | 0.703 | 0.588 | 0.695 | 0.701 | 0.611 | 0.635 | 0.702 | 0.596 | 0.562 | 0.624 | 0.617 | 1.000 | 0.717 |
| AU27=0 | 0.975 | 0.976 | 0.879 | 0.964 | 0.896 | 0.886 | 0.904 | 0.901 | 0.898 | 0.887 | 0.904 | 0.902 | 1.000 | 1.000 |

(b)

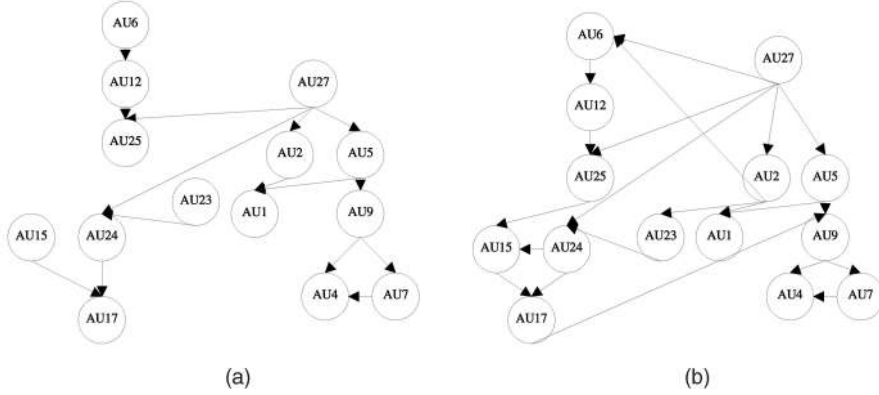


Fig. 4. (a) The prior BN for AU modeling before learning. (b) The learned BN from the training data.

4.3 Learning BN Structure

After analyzing the AU relationships, we obtain an initial BN structure, as shown in Fig. 4a. Although it is our best guess based on the analysis, it may not be correct enough to reflect the true relationships. Therefore, it is necessary to use a large amount of training data to “correct” our “guess” with a structure learning algorithm.

The structure learning algorithm first defines a score that describes the fitness of each possible structure B_s to the observed data, and then, a network structure is identified with the highest score. Suppose we have a domain of discrete variables $\mathbf{X} = \{X_1, \dots, X_N\}$ and a database of sampling data $D = \{D_1, \dots, D_K\}$. Then, the score used for model selection can be defined as

$$\text{Score}(B_s) = \log p(D, B_s) = \log p(D|B_s) + \log p(B_s). \quad (3)$$

The two terms in (3) are actually the log likelihood (the first term) and the log prior probability (the second term) of the structure B_s , respectively. For a large database, the Bayesian information criterion (BIC) [33] introduced by Schwarz has been used to compute the log likelihood $\log p(D|B_s)$ approximately

$$\log p(D|B_s) \approx \log p(D|\hat{\theta}_{B_s}, B_s) - \frac{d}{2} \log(K), \quad (4)$$

where θ_{B_s} is a set of network parameters, $\hat{\theta}_{B_s}$ is the maximum likelihood estimate of θ_{B_s} , d is the number of free parameters in B_s , and K is the number of sampling data in D . Thus, in (4), the first term is used to measure how well the model fits the data, and the second term is a penalty term that punishes the structure complexity.

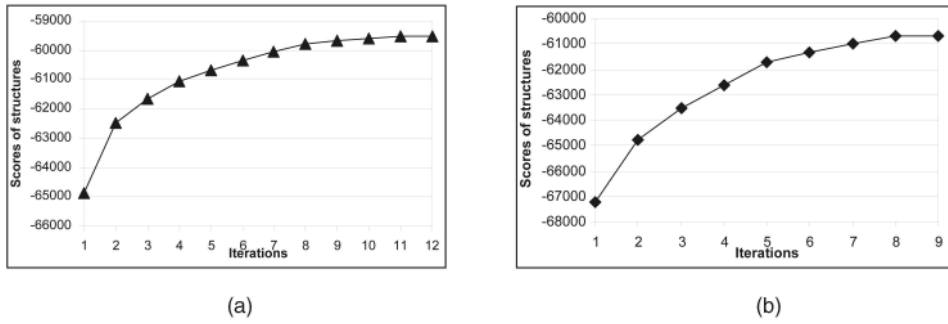


Fig. 5. Scores of structure learning. (a) Starting from a manually constructed network. (b) Starting from a randomly constructed network.

Instead of giving an equal prior $p(B_s)$ to all possible structures, we assign a high probability to the prior structure P_s , which is chosen as the manually constructed model illustrated in Fig. 4a. Let $\Pi_i(B_s)$ be a set of parent nodes of X_i in B_s . Assume that δ_i is the number of nodes in the symmetric difference of $\Pi_i(B_s)$ and $\Pi_i(P_s)$, which is defined as

$$\delta_i = |(\Pi_i(B_s) \cup \Pi_i(P_s)) \setminus (\Pi_i(B_s) \cap \Pi_i(P_s))|, \quad (5)$$

where $\Pi_i(B_s) \cup \Pi_i(P_s)$ represents the union of sets $\Pi_i(B_s)$ and $\Pi_i(P_s)$, $\Pi_i(B_s) \cap \Pi_i(P_s)$ is the intersection of sets $\Pi_i(B_s)$ and $\Pi_i(P_s)$, and “ \setminus ” means set difference.

Then, the prior of any other network ($B_s \neq P_s$) is decreased, depending on the deviation between this network structure and the provided prior network structure [34], as follows:

$$p(B_s) = c\kappa^\delta, \quad (6)$$

where c is a normalization constant to make $\sum_m p(B_{sm}) = 1$, given M possible network structures, $0 < \kappa \leq 1$ is a predefined constant factor, $\delta = \sum_{i=1}^N \delta_i$, and N is the number of nodes in the network.

Having defined a score as in (3), we need to identify a network structure with the highest score by a searching algorithm. To reduce the searching space, a constraint is imposed such that each node has at most two parents for the model simplicity. Thus, the structure learning becomes an optimization problem: find the structure that maximizes the score. In order to avoid getting stuck at a local maximum, an iterated hill climbing is applied in this work:

1. *Initialize the starting network structure.* First, a random network structure (B_s) or the prior structure (P_s) is used as the starting network structure B_0 . In this work, we start with the manually constructed network structure (P_s).
2. *Find a local maximum.* A local search is performed until a local maximum is reached:
 - Starting from B_0 , compute the score of each nearest neighbor of B_0 , which is generated from B_0 by adding, deleting, or reversing a single arc, subject to the acyclicity constraint and the upper bound of parent nodes.
 - Update B_0 with the BN that has the maximum score among all of the nearest neighbors and go back to the previous step until no neighbors have higher score than the current structure.

3. Randomly perturb the current network structure and go back to Step 2 until it converges.

Fig. 5 demonstrates how the score changes when the algorithm starts with the predefined prior structure, shown in Fig. 4a, and a random structure, respectively. Obviously, the final stable score in Fig. 5a is higher than that in Fig. 5b, which is the maximum score obtained by performing several times of the structure learning, starting from the randomly generated network structure. It clearly indicates that the manually constructed structure is a very good starting point to find an appropriate structure that fits the training data well.

The learned structure is shown in Fig. 4b. Basically, the learned structure keeps all of the links in the initial structure and adds several links such as AU27 to AU6, AU2 to AU6, AU25 to AU15, and AU2 to AU23, which are reasonable to reflect facial expressions. For example, AU27 would mostly be contracted in the surprise facial expression, whereas AU6 would mostly happen in the happiness facial expression. The two AUs are seldom present together, which agrees with the co-occurrence table in Table 2a, where $P(AU6 = 1 | AU27 = 1)$ is very small. Overall, the learned structure reflects the data pattern in the training data set better.

4.4 A Dynamic Bayesian Network

In the structure learning, we only focus on the static BN. However, it lacks the ability to express temporal dependencies between the consecutive occurrences of certain AUs in image sequences, whereas AU is an event that develops over time. Furthermore, in addition to the dynamic development of each AU, the relationships among AUs also undergo a temporal evolution. AUs can therefore be better recognized from a sequence of observations over time, instead of from a snapshot. We thus use a dynamic BN to model the dynamic aspect of AUs.

In general, a DBN is made up of interconnected time slices of static BNs, and the relationships between two neighboring time slices are modeled by an HMM such that random variables at time t are influenced by other variables at time t , as well as by the corresponding random variables at time $t - 1$ only. Each time slice is used to represent the snapshot of an evolving temporal process at a time instant. Each slice of the DBN is a static BN, as described in Fig. 4b in Section 4.3, with a measurement node associated with each AU, and the neighboring time slices are interconnected by the arcs joining certain temporal variables from two consecutive slices.

Specifically, in the proposed framework, we consider two types of conditional dependency for variables at two adjacent time slices. In the first type, an arc linking each AU_i node at



Fig. 6. An example image sequence displays the unsynchronized AUs' evolutions in a smile (adapted from [32]).

time $t - 1$ to that at time t depicts how a single AU_i develops over time. In the second type, an arc from AU_i node at time $t - 1$ to AU_j at time t depicts how AU_i at the previous time step affects AU_j ($j \neq i$) at the current time step. In a spontaneous facial behavior, 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, Fig. 6 shows how a smile is developed in an image sequence: first, AU12 is contracted to express a slight smile, and then, AU6 and AU25 are triggered to enhance the happiness. As the intensity of happiness increases, AU12 first reaches its highest intensity level, and then, AU6 and AU25 reach their apices, respectively.

Fig. 7 gives the whole picture of the dynamic BN, including the shaded visual measurement nodes. For presentation clarity, we use the self-arrows to indicate the first type of temporal links as described above. For example, the self-link at AU9 means a temporal link connecting AU9 at time $t - 1$ to AU9 at time t . The arrow from time $t - 1$ to time t indicates the second type of temporal links such as the temporal link from AU12 at time $t - 1$ to AU6 at time t , which are determined manually based on our data analysis.¹

Furthermore, we associate each AU node with a measurement node, which is represented with a shaded circle in Fig. 7. The links between them model the measurement uncertainty that resulted from the computer vision techniques discussed in Section 3. Due to the uncertainty and ambiguity of feature extraction, the conditional dependency between the measurement node and the AU node should be varying rather than fixed, depending on the accuracy and reliability of AU measurement. From the point of view of a BN, the measurement nodes are regarded as observed nodes that provide evidence obtained through some AU recognition techniques such as the computer vision techniques, discussed in Section 3, in the inference procedure. The AU nodes are hidden nodes whose states need to be inferred from the DBN, given the evidence. Here, we should emphasize that the AU measurement extraction and the AU inference are independent; therefore, any AU recognition technique could be employed to obtain the AU measurements. We will discuss the inference procedure later.

Therefore, such a DBN is capable of accounting for uncertainties in the AU recognition process, representing probabilistic relationships among AUs, modeling the dynamics in AU development, and providing principled inference solutions. In addition, with the conditional depen-

ency coded in the graphical model, it can handle situations where several data entries for certain AUs are missing. For example, even if several AU measurements are not available with the computer vision techniques, it is still possible to infer the states of the corresponding AUs by using other available measurements. We will demonstrate it in the experiments.

4.5 Learning DBN Parameters

Given the DBN structure shown in Fig. 7, we need to learn the parameters in order to infer each AU. Learning the parameters in a DBN is actually similar to learning the parameters for a static BN. During DBN learning, we treat the DBN as an expanded BN consisting of two-slice static BNs connected through the temporal variables, as shown in Fig. 7. Besides learning the CPTs for each slice, it is also necessary to learn the transition probabilities between slices. For DBN learning, the training data need to be divided into sets of time sequences of three dimensions representing the sequence index, the node index, and the time slice, respectively. For example, $D_{i,t,l}$ is the value of the i th node in the t th slice of the l th sequence.

Let θ_{ijk} indicate a probability parameter for a DBN with structure B_s as

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

where i ranges over all the variables (nodes in the DBN), j ranges over all the possible parent instantiations for variable X_i , and k ranges over all the instantiations for X_i itself. Therefore, x_i^k represents the k th state of variable X_i , and $pa^j(X_i)$ is the j th configuration of the parent nodes of X_i . For example, a node AU_9^t , shown in Fig. 7, represents the presence/absence of AU9 at time step t , with two binary instantiations [0, 1]. The parents of AU_9^t are AU_5^t , AU_{17}^t , and AU_9^{t-1} , each of which also has two binary instantiations [0, 1]. Hence, there are eight parent configurations for AU_9^t .

Thus, the goal of learning parameters is to maximize the posterior distribution $p(\theta|D, B_s)$ (MAP), given a database D and the structure B_s [35]. Assuming that θ_{ij}^2 are mutually independent [36], then

$$p(\theta|D, B_s) = \prod_{i=1}^n \prod_{j=1}^{q_i} p(\theta_{ij}|D, B_s), \quad (8)$$

where n is the number of variables in the B_s , and q_i is the number of the parent instantiations for X_i . In this work, we use a likelihood-equivalent Bayesian Dirichlet prior probability [35]. Then, the posterior distribution $p(\theta_{ij}|D, B_s)$ can be approximated by Dirichlet distributions

2. $\theta = (\theta_1, \dots, \theta_n)$ is the vector of parameters, where $\theta_i = ((\theta_{ijk})_{k=2}^{q_i})_{j=1}^{q_i}$ are the parameters, and $\theta_{ij} = (\theta_{ij2}, \dots, \theta_{ijq_i})$.

1. Currently, the second type of temporal links is limited to the links from AU_{12}^{t-1} to AU_6^t and from AU_5^{t-1} to AU_9^t since these pairs of AUs have strong temporal relationships based on our data analysis. In the future, we will learn the temporal links among all target AUs from training data automatically.

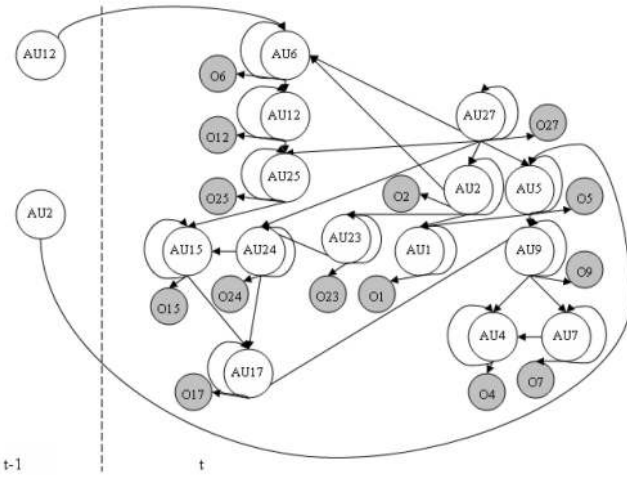


Fig. 7. The DBN for AU modeling. The self-arrow at each AU node indicates the temporal relationship of a single AU from the previous time step to the current time step. The arrow from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the temporal relationship between different AUs. The shaded circle indicates the measurement for each AU.

$$p(\theta_{ij}|D, B_s) = \text{Dir}(\alpha_{ij1} + N_{ij1}, \dots, \alpha_{ijr_i} + N_{ijr_i}), \quad (9)$$

where α_{ijk} reflects the prior belief about how often the case when $X_i = k$ and $pa(X_i) = j$ occurs. N_{ijk} reflects the number of cases in D in which $X_i = k$ and $pa(X_i) = j$, and r_i is the number of all instantiations of X_i .

This learning process considers both prior probability and likelihood so that the posterior probability is maximized. Since the training data set is complete, it can be actually explained as a counting process that results in the following formula of the probability distribution parameters:

$$\theta_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}, \quad (10)$$

where $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

4.6 AU Recognition through DBN Inference

In this section, we describe how the probabilities of different AUs can be inferred, given the AU measurements obtained through the AdaBoost technique, discussed in Section 3.

Dynamic inference estimates the AUs at each time step t . Let AU_i^t indicate the node of AU i at time step t , and $\mathbf{e}^t = \{e_1^t, e_2^t, \dots, e_{27}^t\}$ be the measurements for the 14 AUs at time step t . Thus, the probability of AU_i^t , given the available evidence $\mathbf{e}^{1:t}$, can be inferred by performing the DBN updating process as follows [37]:

- **Prediction.** Given the estimated probability distribution $p(AU_i^{t-1}|\mathbf{e}^{1:t-1})$, which is already inferred at time step $t-1$, we could calculate the predicted probabilities $p(AU_i^t|\mathbf{e}^{1:t-1})$ by using the standard BN inference algorithm such as a version of the junction tree algorithm [38].
- **Rollup.** Remove time slice $t-1$ and use the predictions $p(AU_i^t|\mathbf{e}^{1:t-1})$ for the t slice as the new prior.
- **Estimation.** Add new observations \mathbf{e}^t and calculate the probability distribution over the current state $p(AU_i^t|\mathbf{e}^{1:t})$. Finally, add the slice for $t+1$.

This way, we obtain the posterior probability of each AU, given the observed measurements.

5 EXPERIMENTAL RESULTS

5.1 Facial AU Databases

The dynamic facial AU recognition system is trained on FACS labeled images from two databases. The first database is the Cohn-Kanade *DFAT-504* database [32], which consists of more than 100 subjects covering different races, ages, and genders. This database is collected under controlled illumination and background and has been widely used for evaluating facial AU recognition system. In order to extract the temporal relationships, the Cohn-Kanade database is coded into AU labels frame by frame in our work. Using this database has several advantages: This database demonstrates diversity over the subjects and it involves multiple-AU expressions. The results on the Cohn-Kanade database will be used to compare with other published methods.

However, the image sequences in this database do not contain a complete temporal evolution of the expression, since they only reflect the evolution of the expression starting from a neutral state and ending at the apex, but without the relaxing period, which is a necessary component for a natural facial behavior. Furthermore, this database consists of only 23 expressions, which are only a small part of the enormous facial expressions. Therefore, the relationships among AUs extracted only from this database are not sufficient. Thus, the ISL database is created to overcome the shortcomings of the Cohn-Kanade database.

The ISL database consists of 42 image sequences from 10 subjects displaying facial expressions undergoing a neutral-apex-neutral evolution. The subjects are instructed to perform the single AUs and AU combinations, as well as the six basic expressions. The database is collected under real-world conditions, with uncontrolled illumination and background, as well as moderate head motion. The image sequences are recorded with a frame rate of 30 fps rather than 12 fps, as in the Cohn-Kanade database, and without time stamps occluding some facial features. More importantly, the data are captured while the subject is undergoing a real and natural emotional change, for example, smiling while talking to others and yawning when the subject really felt tired. In addition, the ISL database not only contains the AUs that we intend to detect, but also includes other AUs. Since the proposed system intends to work under real-world situations, where people could display any facial expression besides the six basic expressions, the additional AUs are important in the analysis of spontaneous facial expressions. The ISL database is also coded into AU labels frame by frame manually.

Fig. 8 shows some example images in the ISL database. In Fig. 8a, the subject undergoes the expression state of neutral-laughing-neutral while she smiles (AU6+AU12), opens her mouth (AU25+AU26 (jaw drop)), pitches her head up (AU53 (head up)), looks at another person on the right (AU62 (eyes turn right)+AU63 (eyes up)), and blinks her eyes sometimes (AU45 (blink)). In Fig. 8b, another subject is yawning with a natural head motion, where AU43 (eye closure), AU25, AU27, and AU53 are involved. Both image sequences are not restricted to the basic expressions, as well as the target AUs, and are collected with moderate head/body movement and under natural indoor illumination condition. Therefore, the

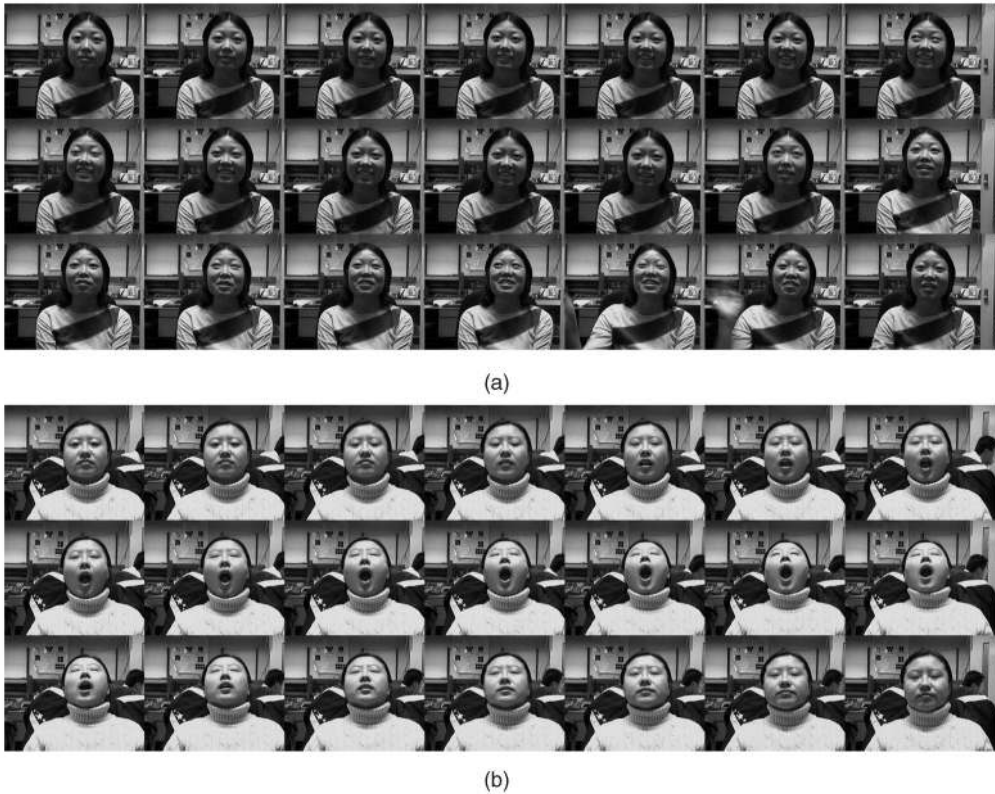


Fig. 8. Some example images from the ISL database. (a) The subject is laughing. (b) The subject is yawning.

results on the ISL database are intended to demonstrate the robustness in real-world applications. Overall, the combined database consists of over 14,000 images from 111 subjects.

Furthermore, the M&M Initiative (MMI) facial expression database collected by Pantic et al. [39] is employed to evaluate the generalization ability of our system. The MMI facial expression database is recorded in true color with a frame rate of 24. The advantage of using this database is that it contains a large number of videos that display facial expressions with a neutral-apex-neutral evolution. Also, the database has been coded into AUs in a frame-by-frame manner.

In the following, we perform validation studies on our system from three aspects: 1) comparison with other state-of-the-art techniques using the Cohn-Kanade database, 2) demonstration of the generalization ability of our system through cross-database validation, and 3) validation of the robustness of our method by testing the system in the real-world environment.

5.2 Evaluation on Cohn-Kanade Database

We first evaluate our system on the Cohn-Kanade database by using leave-one-subject-out cross validation for the 14 target AUs. The AdaBoost classifiers are trained on all of the data, but one subject is left out for testing. For the AdaBoost classifier training, the positive samples are chosen as the training images containing the target AU at different intensity levels, and the negative samples are selected as those images without the target AU regardless the presence of other AUs. The AU labels corresponding to each frame are used for both static and DBN learning. For the BN, we learn the DBN parameters also based on the leave-one-subject-out cross validation.

For a comparison with Bartlett et al.'s method [8], which represents state-of-the-art AU recognition technique, we use the experiment results reported on their Web site [40], which is trained and tested on the Cohn-Kanade database based on the leave-one-subject-out cross validation. Fig. 9 shows the performance for generalization to novel subjects in the Cohn-Kanade database of using the AdaBoost classifiers alone, using the DBN, and using Bartlett et al.'s method, respectively.³ With only the Adaboost classifiers, our system achieves an average recognition rate of 91.2 percent, with a positive rate of 80.7 percent and a false-alarm rate of 7.7 percent for the 14 AUs, where the average rate is defined as the percent of examples recognized correctly. With the use of the DBN, the system achieves the overall average recognition rate of 93.33 percent, with a significant improvement of positive rate of 86.3 percent and improved false-alarm rate of 5.5 percent.

As shown in Fig. 9, for AUs that can be well recognized by the AdaBoost classifier, the improvement by using the DBN is not that significant. However, for the AUs that are difficult to be recognized by the AdaBoost classifier, the improvements are impressive, which exactly demonstrates the benefit of using the DBN. For example, recognizing AU23 (lip tighten) and AU24 is difficult, since the two actions occur rarely, and the appearance changes caused by these actions are relatively subtle. Fortunately, the probability of these two actions' co-occurrence is very high, since they are contracted by the same set of facial muscles. By employing such

3. Ideally, a better comparison can be made through receiver operating characteristic (ROC) analysis. However, we cannot perform an ROC comparison with Bartlett et al.'s method, since they performed ROC analysis by using two databases, one of which is not available to us.

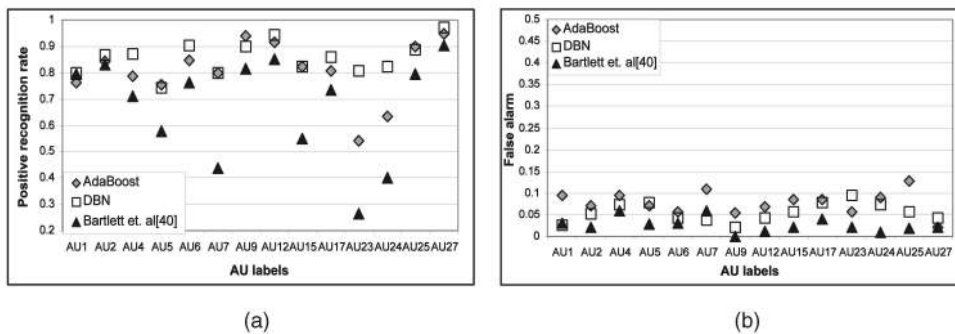


Fig. 9. Comparison of AU recognition results on novel subjects in the Cohn-Kanade database by using the AdaBoost classifier, the DBN, and the results of Bartlett et al. [40], respectively. (a) Average positive rates. (b) Average false-alarm rates.

relationship in the DBN, the positive recognition rate of AU23 increases from 54 percent to 80.6 percent, and that of AU24 increases from 63.4 percent to 82.3 percent. Similarly, by employing the coabsence relationship between AU25 and AU15, the false-alarm rate of AU25 reduces from 13.3 percent to 5.7 percent.

Compared with the system performance reported by Bartlett et al. at their Web site [40] (with an overall average recognition rate of 93.6 percent), our system achieves a similar average recognition rate (93.33 percent). With a marginal increase in the false-alarm rate (about 3 percent), our system significantly increases the positive rate from 70.4 percent (in Bartlett et al's method) to 86.3 percent. More importantly, for some specific AUs that are difficult to be recognized, the improvements are impressive. For example, we obtain an 80.6 percent positive recognition rate of AU23 compared with 26.3 percent by Bartlett et al., and an 82.3 percent positive recognition rate for AU24 compared with 40 percent in their method. In addition, their method only uses the last frame of each sequence that contains the peak AUs, and the first frame of each sequence containing the neutral expression in the database for training and testing, whereas we use the whole image sequences that include the AUs with weak intensity, which are more difficult to recognize. That is why they only use 626 images for testing, whereas we use 8,456 images.

Tian et al. [11] achieve a similar performance by training and testing the upper face AUs on the Ekman-Hager database and the lower face AUs on the Cohn-Kanade database, respectively, where the database is split into the training and testing sets to make the training and testing subjects mutually exclusive. However, manual intervention is required in the initial frame with neutral expression in their method. Valstar et al. [13] report an 84 percent average recognition rate on the Cohn-Kanade database while training on MMI facial expression database [39]. Kapoor et al. [16] obtain an 81.2 percent recognition rate on five upper AUs by using only 25 subjects in the Cohn-Kanade database, with hand-marked pupil positions based on the leave-one-subject-out cross validation. In summary, the average performance of our system is equal to or better than the previously reported systems. However, our system achieved much better performance on some difficult AUs.

To further demonstrate the performance of our system, we performed an ROC analysis for each AU, as shown in Fig. 10. The ROC curve of the AdaBoost is obtained by plotting the true-positive rates against the false-positive rates while varying the decision threshold of AdaBoost. Since the DBN

inference results would be affected by the change from all the AU measurements, we change the decision threshold of AdaBoost for only one target AU while fixing the AdaBoost thresholds for other AUs to obtain the ROC curve for a specific AU. Fig. 10 clearly shows that the area under the ROC curve of the DBN is larger than that of AdaBoost, which further demonstrates that the system performance is improved by using the DBN. Especially for certain AUs (for example, AU23 and AU24) with lower recognition performances by the AdaBoost classifiers, the ROC curves show more significant improvement with the DBN inference.

One significant advantage worth mentioning about the proposed framework is that the use of the DBN tends to always compensate the errors with the measurements from the underlying computer vision techniques. Therefore, if the proposed DBN was built on any of the AU recognition techniques mentioned in the related work, then their results could be improved with low extra cost, given the fact that BN inference is quite efficient in the simple network structure.

5.3 Generalization Validation across Different Databases

In order to evaluate the generalization ability of our system, the system is trained on the Cohn-Kanade database and tested on the MMI facial expression database [39]. The system training is performed using all of the videos from the Cohn-Kanade database. Since most of the videos in the MMI database have only single AU active, we only choose 54 video sequences containing two or more target AUs from 11 different subjects. Fig. 11 shows the testing results of using AdaBoost classifiers alone and using the DBN, respectively, on the MMI facial expression database. With the use of the DBN, the overall average recognition rate is improved from 91.3 percent to 93.9 percent, with an increase in positive rate from 76.5 percent to 78.8 percent and a decrease in the false-alarm rate from 7.3 percent to 4.7 percent. The system still achieves similar recognition performance on novel subjects from a different database. This demonstrates the generalization ability of our system.

5.4 Experiment Results under Real-World Condition

In the third set of the experiments, the system is trained on all of the training images from the Cohn-Kanade database and 38 image sequences with six subjects from the ISL database. The four remaining image sequences with four different subjects from the ISL database are used for testing. This experiment intends to demonstrate the system robustness for real-world environment. The system performance is reported

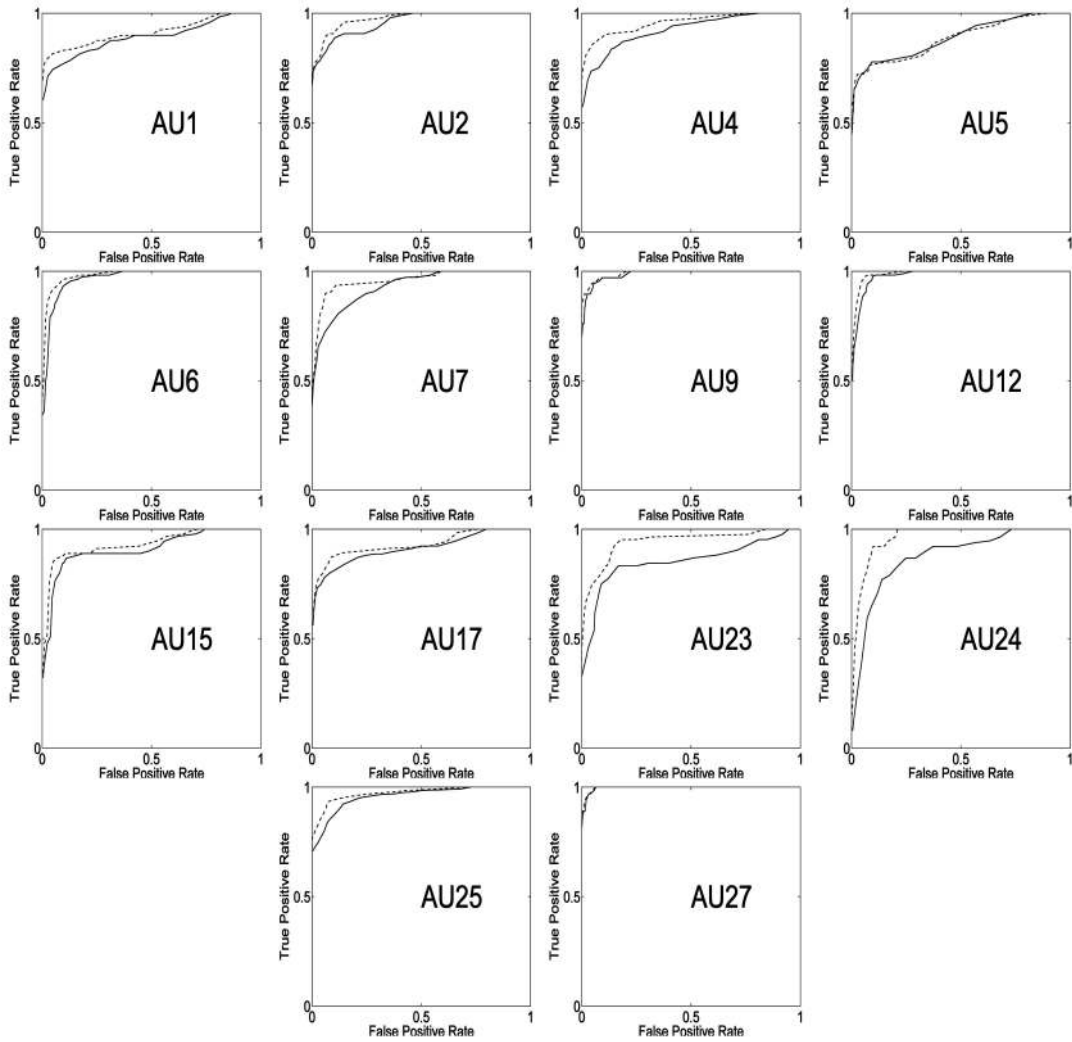


Fig. 10. ROC curves for the 14 target AUs on the Cohn-Kanade database: the solid line represents the ROC curve by AdaBoost, and the dash line represents the ROC curve by the DBN, respectively.

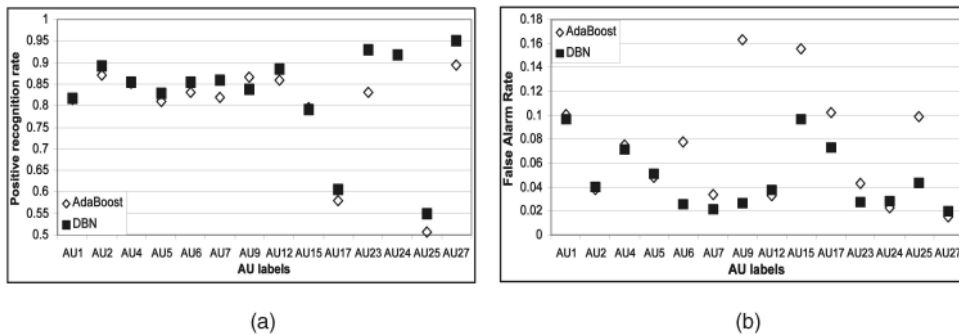


Fig. 11. Generalization performance across different databases. The system is trained on the Cohn-Kanade database and tested on the MMI facial expression database. (a) Average positive rates. (b) Average false-alarm rates.

in Fig. 12. The average recognition rate of the DBN inference is 93.27 percent, with an average positive rate of 80.8 percent and a false-alarm rate of 4.47 percent. Compared with the AU recognition results from the frame-by-frame AdaBoost classification, the AU recognition is improved significantly. The overall correct recognition rate is improved by 5.1 percent, with an 11 percent increase in positive recognition rate and a 4.4 percent decrease in false alarm. Especially for the AUs that are difficult to be recognized, the system

performance is greatly improved. For example, the recognition rate of AU7 (lid tightener) is increased from 84 percent to 94.8 percent, the recognition rate of AU15 is improved from 71.5 percent to 82.9 percent, and that of AU23 is increased from 82.3 percent to 94.4 percent.

The system enhancement comes mainly from two aspects. First, the erroneous AU measurement could be compensated by the relationships among AUs in the BN model. As mentioned above, the AU measurement extracted by the

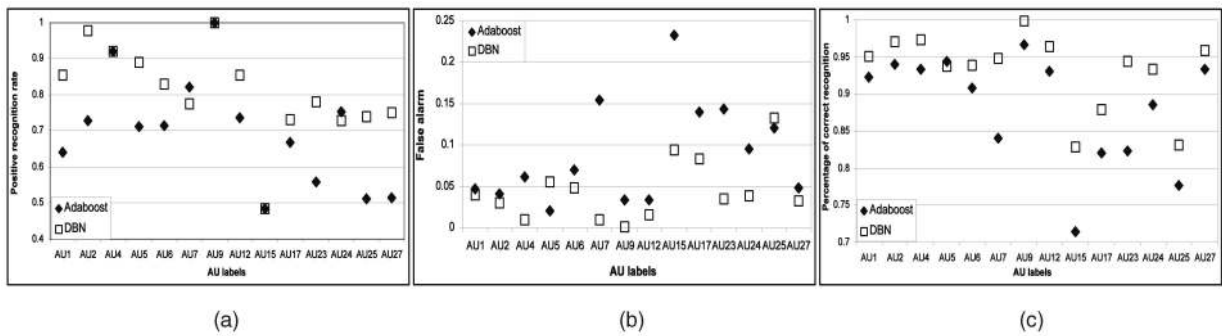


Fig. 12. Comparison of AU recognition results on novel subjects under real-world circumstance by using the AdaBoost classifier and the DBN, respectively. (a) Average positive rates. (b) Average false-alarm rates. (c) Average recognition rates.

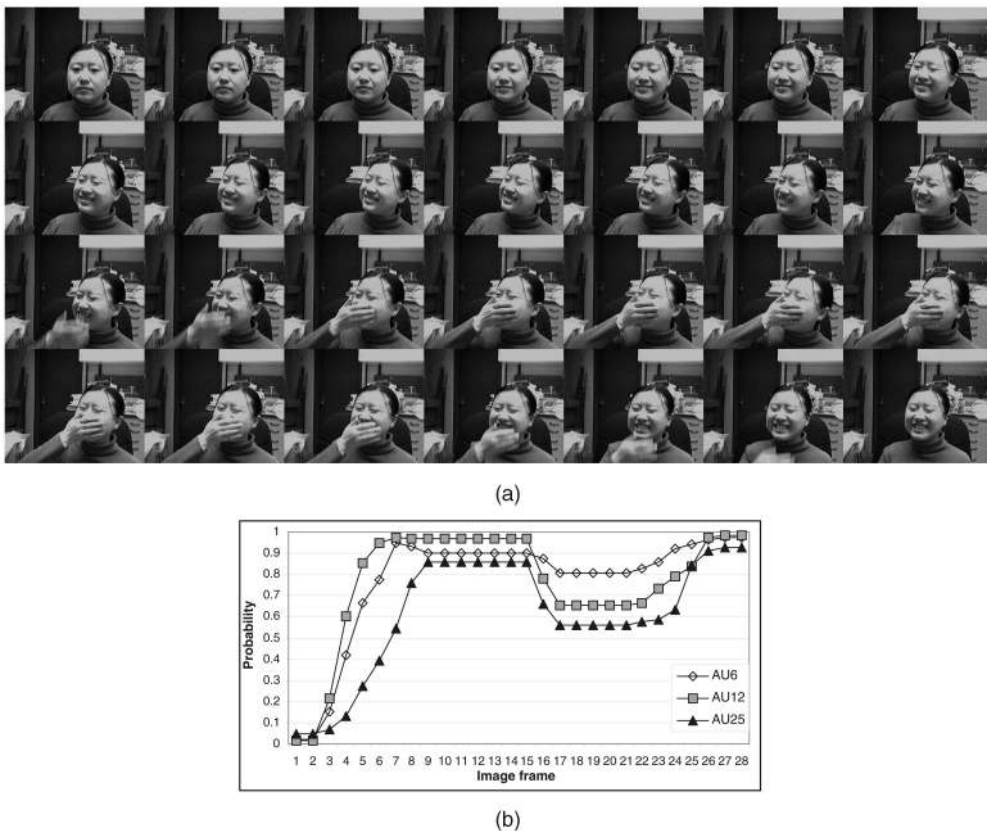


Fig. 13. (a) An image sequence where the face undergoes an out-of-image plane rotation and the mouth is occluded by hand in some frames. (b) The x-axis represents the image frame number and the y-axis represents the probabilities of the involved AUs in the corresponding frames. The edges between the points depict the temporal evolution of each AU. Occlusion occurs from frame 15 to frame 25.

AdaBoost classification is sensitive to inaccurate image alignment. Moreover, some AUs are inherently difficult to be identified. For example, the recognition of AU7 is difficult, since the contraction of AU7 would narrow the eye aperture, which, however, could also be caused by AU6 or by a relaxed eyelid. Fortunately, AU7 occurs often with AU4, which can be easily recognized. By embedding such relationship in the DBN, the false-alarm rate of AU7 is reduced greatly (from 15.4 percent to 1 percent).

Second, and, more importantly, the proposed approach could recognize an individual AU even if there is no direct measurement for that AU due to face pose variation or occlusion. Fig. 13a gives an image sequence where the face pose varies in some frames. Since the AdaBoost classifiers are trained on nearly frontal view faces in our current work,

although the on-image-plane head rotation and translation could be compensated by image normalization given the eye positions, some of the AU measurements obtained by the AdaBoost classifiers may not be accurate under large out-of-image-plane head rotation. In the proposed system, the inaccurate AU measurements could be compensated by their relationships with other AUs through the DBN inference, as shown in Fig. 13b.

Fig. 13a also shows that the mouth is occluded in some frames due to occlusion by hand. The AdaBoost classifier could not recognize the AUs in the lower face region due to the occlusion. Fig. 13b depicts the probabilities of the involved AUs by the proposed approach for the image sequence. In Fig. 13b, we could observe that although there are no direct measurements for AU12 and AU25 from frame

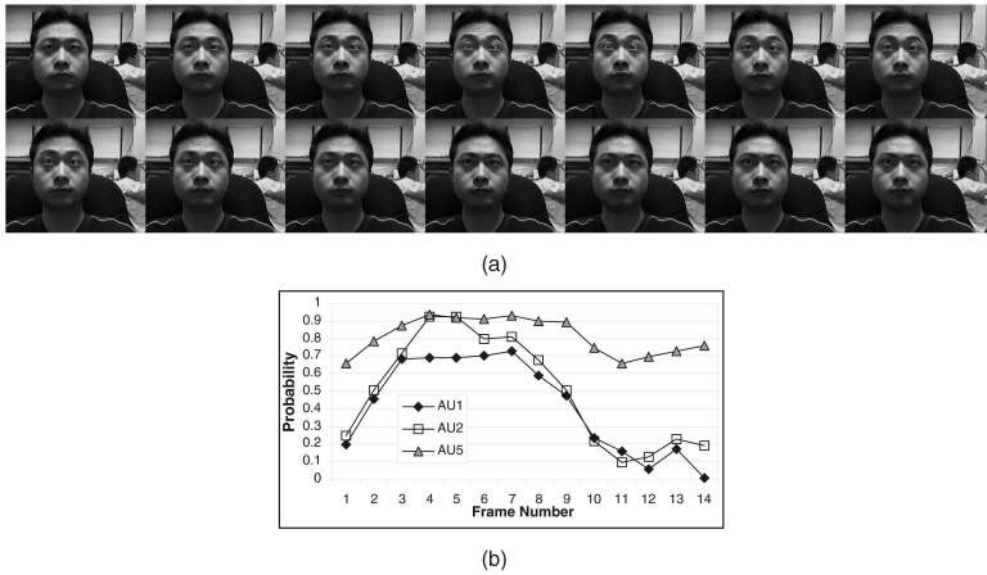


Fig. 14. Automatic AU recognition on image sequence under a real-world circumstance. (a) A subject with eyebrow rising and eye lid rising in an image sequence. (b) The probabilities of the involved AUs in the corresponding frames, where each curve represents the temporal evolution of a specific AU, respectively.

15 to frame 25, they could still be correctly estimated through their semantic relationships with the other AUs, as well as the temporal evolution of themselves, since $AU6+AU12+AU25$ implies a happy expression.

Fig. 14 shows the system output (the probabilities of AUs) for an image sequence involving multiple facial actions. It can be seen that the system output changes smoothly, following the temporal evolution of the facial actions over time. Since there are individual differences with respect to the magnitudes of AUs, it is difficult to determine the absolute intensity of a given subject. Our dynamic modeling of facial AUs can more realistically reflect the evolution of a spontaneous facial emotion and, thus, can extract the relative intensity changes of the AUs.

6 CONCLUSIONS AND FUTURE WORK

In this paper, a fully automatic system is proposed for real-time recognition of AUs in real-world environment. Instead of recognizing each AU or AU combination individually or statically as other people usually do, we employ a DBN to model both the semantic and temporal relationships among various AUs. Such a model is capable of representing the relationships among different AUs in a coherent and unified hierarchical structure, accounting for the nonuniqueness (that is, the nonadditive effect) and uncertainties in the AU recognition process, modeling the dynamics in facial behavior development and providing principled inference. Especially, both the structure and parameters are learned with advanced machine learning techniques to systematically combine the domain information and statistical information extracted from the data. Therefore, under such a framework, the erroneous AU measurement from the underlying computer vision techniques could be compensated by exploiting the relationships among AUs: some AUs that are difficult to be directly recognized can be inferred indirectly from other AUs, and the low-intensity AUs could be recognized more reliably by using the information from

other high-intensity AUs. As shown in the experiments, the DBN framework integrated with the underlying feature extraction techniques yields significant improvement of AU recognition over using computer vision techniques alone. Compared with the state-of-the-art methods, our system achieves impressive improvements on the publicly available database, especially for the AUs that are difficult to be recognized. The improvement is even more obvious under real-world environment such as illumination variation, face pose variation, and occlusion. Furthermore, our system is also demonstrated to generalize to different databases.

Currently, the full automatic AU recognition system can process about 7 fps in 320×240 images on a 2.8-GHz Pentium 4 PC. The system could be speeded up by employing a fast Gabor filtering algorithm [41].

In this work, we have learned the strong and common AU relationships from two databases. However, if the learned DBN model is used for a very different application, where some AU relationships are not represented in the training data, then we should relearn the model from the new training data in order to better characterize the AU relationships in the target application.

In the current framework, we do not recognize the absolute intensities of AUs, and we only focus on the 14 most common AUs. We plan to study more AUs and recognize the different levels of AU intensity in the future by further improving the computer vision techniques. Moreover, introducing more AUs and AU intensities would be more challenging for the DBN modeling procedure and learning approach since AUs need to be represented at finer granularity. In addition, how we can obtain the ground truth for AUs at different levels, especially for AUs at low-intensity levels, represents another challenge. This would be another future work. Finally, it would be also interesting to identify facial events based on the AU recognition results. In the long run, we would like to apply the proposed techniques to human emotion recognition, human expression animation, and other human-computer interaction applications.

ACKNOWLEDGMENTS

Portions of the research in this paper use the MMI Facial Expression Database collected by M. Pantic and M.F. Valstar. We also use the Cohn-Kanade *DFAT-504* database. We gracefully acknowledge their support. This project is supported in part by the US Air Force Office of Scientific Research (AFOSR) Grant F49620-03-1-0160 and the DARPA/US Office of Naval Research (ONR) Grant N00014-03-1-1003.

REFERENCES

- [1] P. Ekman and W.V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
- [2] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge Univ. Press, 1982.
- [3] J.J. Lien, T. Kanade, J.F. Cohn, and C. Li, "Detection, Tracking, and Classification of Action Units in Facial Expression," *J. Robotics and Autonomous System*, vol. 31, pp. 131-146, 2000.
- [4] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct. 1999.
- [5] B. Fasel and J. Luetttin, "Recognition of Asymmetric Facial Action Unit Activities and Intensities," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 1100-1103, 2000.
- [6] E. Smith, M.S. Bartlett, and J.R. Movellan, "Computer Recognition of Facial Actions: A Study of Co-Articulation Effects," *Proc. Eighth Ann. Joint Symp. Neural Computation*, 2001.
- [7] J.J. Bazzo and M.V. Lamar, "Recognizing Facial Actions Using Gabor Wavelets with Neutral Face Average Difference," *Proc. Sixth IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 505-510, 2004.
- [8] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and R. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 568-573, 2005.
- [9] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang, "Facial Expression Recognition from Video Sequences: Temporal and Static Modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160-187, 2003.
- [10] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics*, vol. 1, pp. 610-616, 2004.
- [11] Y. 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.
- [12] Y. Tian, T. Kanade, and J.F. Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity," *Proc. Fifth IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 218-223, 2002.
- [13] M.F. Valstar, I. Patras, and M. Pantic, "Facial Action Unit Detection Using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, Workshop Vision for Human-Computer Interaction*, 2005.
- [14] Y. Zhang and Q. Ji, "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 699-714, May 2005.
- [15] H. Gu and Q. Ji, "Facial Event Classification with Task-Oriented Dynamic Bayesian Network," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 870-875, 2004.
- [16] A. Kapoor, Y. Qi, and R.W. Picard, "Fully Automatic Upper Facial Action Recognition," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pp. 195-202, 2003.
- [17] R. El Kaliouby and P.K. Robinson, "Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2004.
- [18] M. Pantic and L.J.M. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images," *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 34, no. 3, pp. 1449-1461, June 2004.
- [19] I. Cohen, F.G. Cozman, N. Sebe, M.C. Cirelo, and T.S. Huang, "Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553-1567, Dec. 2004.
- [20] P. Ekman, W.V. Friesen, and J.C. Hager, "Facial Action Coding System: The Manual," Research Nexus Division, Network Information Research Corp., Salt Lake City, 2002.
- [21] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743-756, July 1997.
- [22] J.F. Cohn and A. Zlochower, "A Computerized Analysis of Facial Expression: Feasibility of Automated Discrimination," *Am. Psychological Soc.*, 1995.
- [23] M. Valstar, M. Pantic, and I. Patras, "Motion History for Facial Action Detection in Video," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics*, vol. 1, pp. 635-640, 2004.
- [24] C. Huang and Y. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," *J. Visual Comm. and Image Representation*, vol. 8, no. 3, pp. 278-290, 1997.
- [25] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," *J. Personality and Social Psychology*, vol. 37, no. 11, pp. 2049-2058, 1979.
- [26] 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.
- [27] P. Viola and M. Jones, "Robust Real-Time Object Detection," *Int'l J. Computer Vision*, vol. 57, no. 2, pp. 137-154, May 2004.
- [28] P. Wang, M.B. Green, Q. Ji, and J. Wayman, "Automatic Eye Detection and Its Validation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, Workshop Face Recognition Grand Challenge Experiments*, vol. 3, 2005.
- [29] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 947-954, 2005.
- [30] T. Lee, "Image Representation Using 2D Gabor Wavelets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959-971, Oct. 1996.
- [31] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," *Proc. Third IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 454-459, 1998.
- [32] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. Fourth IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 46-53, 2000.
- [33] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, pp. 461-464, 1978.
- [34] D. Heckerman, D. Geiger, and D.M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, vol. 20, no. 3, pp. 197-243, 1995.
- [35] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," Technical Report MSR-TR-95-06, Microsoft Research, 1995.
- [36] D. Spiegelhalter and S. Lauritzen, "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks*, vol. 20, pp. 579-605, 1990.
- [37] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*. Chapman and Hall/CRC, 2004.
- [38] U. Kjaerulff, "DHUGIN: A Computational System for Dynamic Time-Sliced Bayesian Networks," *Int'l J. Forecasting-Special Issue on Probability Forecasting*, vol. 11, pp. 89-111, 1995.
- [39] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo*, July 2005.
- [40] M. Stewart Bartlett, G. Littlewort, J. Movellan, and M.S. Frank, "Auto FACS Coding," <http://mplab.ucsd.edu/grants/project1/research/fully-auto-facs-coding.html>, 2007.
- [41] I. Young, L. van Vliet, and M. van Ginkel, "Recursive Gabor Filtering," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2798-2805, 2002.



Yan Tong received the BS degree from Zhejiang University, China, in 1997 and the MS degree from the University of Nevada, Reno, in 2004. She is currently pursuing the PhD degree at Rensselaer Polytechnic Institute, Troy, New York. Her areas of research include computer vision, pattern recognition, and human computer interaction. She is a student member of the IEEE.



Wenhui Liao received the PhD degree from the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York, in 2006. Her areas of research include probabilistic graphical models, information fusion, computer vision, and human-computer interaction. She is a member of the IEEE.



Qiang Ji received the PhD degree in electrical engineering from the University of Washington in 1998. He is currently an associate professor in the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI), Troy, New York. Prior to joining RPI in 2001, he was an assistant professor in the Department of Computer Science, University of Nevada, Reno. He also held research and visiting positions with Carnegie Mellon University, Western Research, and the US Air Force Research Laboratory. His research interests include computer vision, probabilistic reasoning with Bayesian networks for decision making and information fusion, human-computer interaction, pattern recognition, and robotics. He has published more than 100 papers in peer-reviewed journals and conferences. His research has been funded by local and federal government agencies including the US National Science Foundation (NSF), the US National Institute of Health (NIH), the US Air Force Office of Scientific Research (AFOSR), the US Office of Naval Research (ONR), DARPA, and the US Army Research Office (ARO) and by private companies including Boeing and Honda. He is a senior member of the IEEE.

► **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.**