Measuring the intensity of spontaneous facial action units with dynamic Bayesian network

Yongqiang Li a,*, S. Mohammad Mavadati b, Mohammad H. Mahoor b, Yongping Zhao a, Qiang Ji c

a School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China
b Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80208, USA
c Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

A R T I C L E   I N F O

Article history:
Received 27 March 2014
Received in revised form 16 March 2015
Accepted 20 April 2015
Available online 29 April 2015

Keywords:
Spontaneous facial expression
AU intensity
Dynamic Bayesian Network
FACS
DISFA database

A B S T R A C T

Automatic facial expression analysis has received great attention in different applications over the last two decades. Facial Action Coding System (FACS), which describes all possible facial expressions based on a set of facial muscle movements called Action Unit (AU), has been used extensively to model and analyze facial expressions. FACS describes methods for coding the intensity of AUs, and AU intensity measurement is important in some studies in behavioral science and developmental psychology. However, in majority of the existing studies in the area of facial expression recognition, the focus has been on basic expression recognition or facial action unit detection. There are very few investigations on measuring the intensity of spontaneous facial actions. In addition, the few studies on AU intensity recognition usually try to measure the intensity of facial actions statically and individually, ignoring the dependencies among multilevel AU intensities as well as the temporal information. However, these spatiotemporal interactions among facial actions are crucial for understanding and analyzing spontaneous facial expressions, since these coherent, coordinated, and synchronized interactions are that produce a meaningful facial display. In this paper, we propose a framework based on Dynamic Bayesian Network (DBN) to systematically model the dynamic and semantic relationships among multilevel AU intensities. Given the extracted image observations, the AU intensity recognition is accomplished through probabilistic inference by systematically integrating the image observations with the proposed DBN model. Experiments on Denver Intensity of Spontaneous Facial Action (DISFA) database demonstrate the superiority of our method over single image-driven methods in AU intensity measurement.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Facial expression is one of the most common nonverbal communication media that individuals use in their daily social interactions. Analyzing facial expression will provide powerful information to describe the emotional states and psychological patterns of individuals. In the last two decades automatic facial expression recognition has gained more attention in several applications in developmental psychology, social robotics, affective online tutoring environment and intelligent Human–Computer Interaction (HCI) design [1,2].

Facial Action Coding System (FACS) is one of the well-known approaches for describing and analyzing facial expressions [3]. FACS describes all possible facial expressions based on a set of anatomical facial muscle movements, called Action Unit (AU). For instance, AU12 or lip corner puller specifies contractions that occur on the face by Orbicularis oculi muscle [3]. FACS is also capable of representing the dynamics of every facial behavior by annotating the intensity of each AU in a five ordinal scale (i.e. scales A–E that indicate the barely visible to maximum intensity of each AU). AU intensity can describe the occurrence of spontaneous facial expressions in more detail. The general relationship between the scale of evidence and the A–B–C–D–E intensity scoring, as well as some AU samples is illustrated in Fig. 1. Generally, the A level refers to a trace of the action; B, slight evidence; C, marked or pronounced; D, severe or extreme; and E, maximum evidence. For example, “AU12B” indicates AU12 with a B intensity level. Manual FACS coding is an intensive and time consuming task and designing an automatic system which can specify the list of AUs and their intensities would help the community to analyze spontaneous facial behaviors accurately and efficiently.

* Corresponding author. Tel.: +86 15243504277.
E-mail address: liyongqiang@hit.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.patcog.2015.04.022
0031-3203/© 2015 Elsevier Ltd. All rights reserved.
Majority of the existing literature has been focusing on two types of facial expression studies. The first category is concerned with analyzing and classifying prototypic facial expressions (aka as six basic expressions: happy, sad, disgust, anger, surprise, fear). These studies are mostly designed to recognize the basic expressions that can represent the human emotions. These expressions of emotions are known to be similar among different cultures [39]. The second category of facial behavior analysis specifies expressions by a set of AUs where the goal is to represent and recognize facial AUs defined by FACS. The latter approach can comprehensively describe a wider range of facial expressions. AU-based analyzers are also capable for representing the prototypic facial expressions as a combination of AUs. For instance ‘fear’ can be represented by combination of AU1, 2, 4, 5, and 25 [3].

Most of the existing studies have been focused on prototypic facial expressions and detecting the occurrence of AUs in posed facial expressions. In many real-world applications, we need to analyze spontaneous facial expression, such as categorizing pain related facial expressions [7] and measuring the engagement of students for online tutoring applications. Spontaneous facial behavior analysis can be very challenging due to several factors, such as out-of-plane head motion and different poses, subtle facial expressions and intra-subject variability in dynamics and timing of different facial actions. In addition it has been shown that the dynamics and patterns of spontaneous facial expressions can be very different from the posed ones.

Analyzing spontaneous facial expressions, especially for intensity measurement, is not as robust and accurate as the posed one, because of the aforementioned challenging factors. In early works for automatic AU intensity measurement, Bartlett et al. [13] measured the intensity of AUs in posed and spontaneous facial expressions by using Gabor wavelet and support vector machines. The mean correlation with human-coded intensity of their automated face recognition system for posed and spontaneous facial behavior is 0.63 and 0.3, respectively. These quantitative results demonstrate that recognizing spontaneous expressions is more challenging than posed expressions.

In the area of spontaneous facial actions recognition, there are very few works on detecting or measuring the intensity of spontaneous facial actions [5,15]. To the best of the authors’ knowledge, most of the current studies, including [5,15,14], analyze spontaneous facial actions statically and individually. In other words, the dependencies among multilevel AU intensities as well as the temporal information are ignored. The semantic and dynamic relationships among facial actions are crucial for understanding and analyzing spontaneous expression. In fact, the coordination and synchronized spatiotemporal interactions between facial actions produce a meaningful facial expression. Tong et al. [25] employed dynamic Bayesian Network (DBN) to model the dependencies among AUs and achieved improvement over single image-driven methods, especially for recognizing AUs that are difficult to detect but have strong relationships with other AUs. However, their work [25] focuses on AU detection of posed expression.

Following the idea in [25], in this paper, we introduce a framework based on DBN to systematically model the spatiotemporal dependencies among multi-AU intensity levels in multiple frames, in order to measure the intensity of spontaneous facial actions. The proposed probabilistic framework is capable of recognizing multi-level AU intensities in spontaneous facial expressions. Denver Intensity of Spontaneous Facial Action (DISFA) database [9] is employed in this study. DISFA database is publicly available for analyzing the AU intensities and their dynamics. For every frame in DISFA, the intensity of every AU within the scale of 0 (absence of an AU) to 5 (maximum intensity) has been provided. To demonstrate the effectiveness of the proposed model, rigorous experiments are performed on DISFA database. The experimental results as well as the detailed analysis on the improvements are reported in this paper.

2. Related works

Given the significant role of faces in human’s emotional and social life, automating the analysis of facial expression has gained great attention in both academia and industry. An automated facial expression recognition system usually consists of two key stages: feature extraction and machine learning algorithm design for classification. Commonly used features that represent facial gestures or facial movements include optical flow [35,41], explicit feature measurement (e.g., length of wrinkles and degree of eye opening) [42], Haar features [43], Local Binary Patterns (LBP) features [44,45], independent component analysis (ICA) [46], feature points [36], 3D geometries features [54,55], and Gabor wavelets [4]. Given the extracted facial features, facial actions are identified using machine learning methods such as Neural Networks [42], Support Vector Machines (SVM) [5], rule-based approaches [47], AdaBoost classifiers, and Sparse Representation (SR) classifiers [48].

The recent attempt of facial action intensity estimation has received increasing interest due to the semantic richness of the predictions. Bartlett et al. [49] estimated the intensity of action units by using the distance of a test example to the SVM separating hyperplane, while Hamm et al. [50] used the confidence of the decision obtained from AdaBoost. Multi-class classifier is a natural choice for this problem. For example, Mahoor et al. [5] utilized AAM features in conjunction with six one-vs-all SVM classifiers to automatically measure the intensity of AU6 and AU12 in videos captured from infant–mother communications. More recently, Mavadati et al. presented the DISFA database in [9], which is fully coded by six level AU intensity and employed Local Binary Pattern histogram (LBPH) features, HOG features, and Gabor features followed by multi-class SVM classification for AU intensity measurement. Besides multi-class classifiers, regressor is also commonly used for the task of continuous expression estimation. For instance, Jeni et al. [51,56] and Savran et al. [52] applied Support Vector Regression (SVR) for prediction, while Kaltwang et al. [53] used Relevance Vector Regression (RVR) instead. Both methods, SVR and RVR, are extensions to regression of SVM, although RVR yields a probabilistic output.

Most of the current works recognize facial actions individually and statically. However, due to the richness, ambiguity, and
dynamic nature of facial actions, individually recognizing each AU intensity is not accurate and reliable for spontaneous facial expressions. Understanding spontaneous facial expression requires not only improving facial motion observations, but more importantly, exploiting the spatiotemporal interactions among facial motions since the coherent, coordinated, and synchronized interactions among AUs produce a meaningful facial expression.

Some previous studies focused on exploiting the semantic and dynamic relationships among facial actions [35,35,25]. Lien et al. [35] and Valstar et al. [36] employed a set of Hidden Markov Models (HMMs) to represent the facial actions evolution in time. Tong et al. [25,37] constructed a Dynamic Bayesian Network (DBN) based model to further exploit the semantic and temporal dependencies among facial actions. However, these works are all limited to detection of the presence and absence of AUs mainly in posed expressions. Recently, there are some works exploiting the interactions between AU intensity values. Sandbach et al. [58] employed a Markov Random Field to model the static correlation relationship among AU intensities, which improves the recognition accuracy compared to regressors, i.e., SVRs and Directed Acyclic Graph Support Vector Machines. A Conditional Ordinal Random Fields (CORF) is extended [59] to application in AU intensity estimation. However, this method [59] can only estimate the AU intensity when the presence of the AU is already known. Baltrušaitis et al. [57] combine ANNs and Continuous Conditional Random Fields (CCRF) as a Continuous Conditional Neural Fields (CCNF) for structured regression for AU intensity estimation, where each hidden component consists of a neutral layer.

In this work, we construct a DBN to systematically model the spatiotemporal interactions among multi-level AU intensities. Advanced machine learning techniques are employed to train the framework from both subjective prior knowledge and training data. The proposed method differs from previous works [25,37] in both theory and applications. Theoretically, this paper focuses on exploiting the AU intensity correlations, and modeling the spatiotemporal dependencies among multi-level AU intensities. In terms of applications, the focus of this paper is to design and develop an automatic system to measure the intensity of spontaneous facial action units, which is much more challenging than detecting posed facial action units.

Fig. 2 gives the flowchart of the proposed online AU intensity recognition system. After registering facial images, we utilized two well-known feature extraction techniques that are capable of representing the appearance information. These features are Histogram of Oriented Gradient (HOG), and Localized Gabor Features which are described below.

3. AU intensity observation extraction

In this section we describe our AU intensity image observation extraction method, which consists of face registration, facial image representation, dimensionality reduction and SVM classification.

3.1. Face registration

Image registration is a commonly used technique to align similar data (i.e. the reference and sensed image). In order to register two images, oftentimes, a set of points called landmark points is exploited for representing and aligning images. In our study, we used 66 landmark points of DISFA database to represent the location of important facial components [9]. To obtain the reference landmark points we averaged the 66 landmark points over the entire training set. A 2D similarity transformation and the bilinear interpolation technique were utilized to transform the new image into the reference coordinate system. The registered images are then masked to extract the facial regions and resized to 128 × 108 pixels.

3.2. Facial image representation

After registering facial images, we utilized two well-known feature extraction techniques that are capable of representing the appearance information. These features are Histogram of Oriented Gradient (HOG), and Localized Gabor Features which are described below.
3.2.1. Histogram of oriented gradient

The histogram of oriented gradient was firstly applied in human detection area [16], however it has been reported to be a powerful descriptor for representing and analyzing facial images [9]. HOG is a descriptor which counts the occurrences of gradient orientation in localized portion of an image and it can efficiently describe the local shape and appearance of an object. To represent the spatial information of an object, images are divided into small cells and for each cell, the histogram of gradient is calculated. (For more information about gradient filters and number of histogram bins, interested readers are referred to [16].)

In our experiment, for every image (i.e image size is 128 x 108 pixels), we built a cell with 18 x 16 pixels and in overall 48 cells are constructed out of each image. We applied the horizontal gradient filter [−1 0 1] with 59 orientation bins in our experiments. To construct the HOG feature vector, the HOG representation for all the cells was concatenated together and finally the HOG feature vector with size 2832 (48 x 59) was obtained.

3.2.2. Localized Gabor features

Gabor wavelet is another famous technique for representing texture information of an object. A Gabor filter is defined with a Gaussian kernel that is modulated by a sinusoidal plane. Gabor feature has a powerful capability for representing facial textures and has been used for different applications including facial expression recognition [17]. In our experiment, to efficiently extract both texture and shape information of facial images, 40 Gabor filters (i.e. 5 scales and 8 orientations) were applied to regions defined around every 66 landmark points and as a result 2640 Gabor features were extracted.

3.3. Dimensionality reduction: manifold learning

In many real world applications in machine learning and pattern classification, high-dimensional features make analyzing the samples more complicated, e.g. 2832 dimensional HOG features and 2640 dimensional Gabor features in this work which contain redundant information. Reviewing the literatures has shown that facial expression appearances are embedded along a low dimensional manifold in high-dimensional space [19,20]. Therefore employing a non-linear technique (i.e., manifold learning), which can preserve the local information of facial appearances can be an effective approach to represent and classify facial expressions. Manifold learning is a nonlinear technique, which assumes that the data points are sampled from a low dimensional manifold but they are embedded into a high dimensional space. Mathematically speaking, given a set of points $x_1, ..., x_n \in \mathbb{R}^d$ find a set of point $y_1, ..., y_n \in \mathbb{R}^{(d \setminus D)}$ such that $y_i$ represents $x_i$ efficiently. In this paper we utilized the Laplacian eigenmap algorithm [23] to reduce the dimensionality of data, more specifically, the high-dimensional features were mapped to a space with 29 dimensions. Belkin and Niyogi introduced the Laplacian Eigenmap algorithm [23]. This algorithm aims to map the close points in high dimensional space into the close points in the low dimensional one. The generalized eigenvector problem is applied for solving this problem and the first $d$ eigenvectors corresponding to the first $d$ smallest eigenvalues are used to describe the embedded $d$-dimensional Euclidean space. For more details see [23]. Similar to [5], we utilized Spectral Regression (SR) algorithm to find a projection function which can map the high dimensional data, such as HOG and Gabor features, into low dimensional space.

3.4. Classification

Given the reduced feature vectors, we extract the AU intensity observation through SVM classification [24]. The SVM classifiers aims to find discriminative hyperplane with the maximum margin for dividing data belongs to different classes. There are several parameters, such as kernel’s type (e.g. Linear, Polynomial, Radial Basis Function (RBF) kernels) that can affect the efficiency of the SVM classifier. For more detailed information on SVM read [24]. In our experiment, to classify $C = 6$ AU intensity levels of each AU, we used one-against-one strategy, in which $C(C-1)/2$ binary discriminant functions are defined, one for every possible pair of classes. And we did the same for all 12 AUs of DISFA database, which results in $15 \times 12 = 180$ binary classifiers. We examined three different kernels (i.e. Linear, Polynomial and Gaussian RBF kernels) where the Gaussian RBF outperformed the other two kernels. Although we can extract AU intensities up to some accuracy, this image-appearance-based approach treats each AU and each frame individually and largely relies on the accuracy of face region alignment. In order to model the dynamics of AUs, as well as their semantic relationships. In addition to deal with the image uncertainty, we utilize a DBN for AU inference. Consequently, the output of the SVM classifier is used as the evidence for the subsequent AU inference via the DBN.

4. DBN model for facial action unit intensity recognition

4.1. AU intensity correlation learning

4.1.1. AU intensity correlation analysis

Measuring the intensity of AUs in a single frame of video is difficult due to the variety, ambiguity, and dynamic nature of facial actions. This is especially true for spontaneous facial expressions.

![Fig. 3. Nonadditive effect in an AU combination. (a) AU12 occurs alone. (b) AU15 occurs alone. (c) AU12 and AU15 appear together. (Adapted from [3]).](image-url)
Moreover, when AUs occur in a combination, they may be non-additive, which means that the appearance of an AU in a combination is different from its stand-alone appearance. Fig. 3 demonstrates an example of the non-additive effect: when AU12 (lip corner puller) appears alone, the lip corners are pulled up toward the cheekbone; however, if AU15 (lip corner depressor) is also becoming active, then the lip corners are somewhat angled down due to the presence of AU15. The non-additive effect increases the difficulty of recognizing AU individually.

As described in the FACS manual [3], the inherent relationships among AUs can provide required information to better analyze the facial expressions. The inherent relationships can be summarized as co-occurrence relationships and mutual exclusive relationships. The co-occurrence relationships characterize the groups of AUs, which oftentimes appear together to show meaningful facial emotions. For instance, AU6 + AU12 + AU25 represents 'happiness' (Fig. 4). AU4 + AU15 + AU17 shows 'sadness' (Fig. 4), and AU9 + AU17 is an illustration of 'disgust' (Fig. 4).

On the other hand, considering the alternative rules provided in the FACS manual [3], some AUs are mutually exclusive which means that some AUs rarely happen together in spontaneous expressions in real life. For instance, it may not be possible to demonstrate AU25 (lips part) with AU24 (lip pressor) simultaneously as described in the FACS manual "the lips cannot be parted and pressed together" [3]. Note that "mutual exclusive relationship" in the paper does not mean that the involved AUs can never be present together, but that the involved AUs are exclusive and the probability of the involved AUs occurring simultaneously is pretty low.

In addition to the co-occurrence and mutual exclusive relationships of AUs, there are also some limitations and restrictions on the intensity level of co-occurring action units. For instance, when AU6 (cheek raiser) and AU12 (lip corner puller) are occurring together, as shown in Fig. 4, the high/low intensity of one AU indicates a high probability of high/low intensity of another one. Similarly, when AU10 (upper lip raiser) and AU12 are activated together, if AU12 is having strong intensity, the intensity of AU10 will be forced to be low as described in the FACS manual "such strong actions of AU12 counteract the influence of AU10 on the shape of the upper lip" [3].

Modeling such dependencies among AUs has been proven helpful for increasing the AU recognition accuracy [35,36,25]. For example, Tong et al. [25,37] employed a DBN to systematically model the spatiotemporal relationships among AUs, and achieved a marked improvement over the image observation. However, [25,37] focuses on AU detection, which only recognizes AU's absence or presence. In addition, [25] detects AUs from posed expressions, which are created by asking subjects to deliberately make specific facial actions or expressions. Spontaneous expressions, on the other hand, typically occur in uncontrolled conditions, and are more challenging to measure [13] because of head pose variations, occurrence of subtle facial behaviors. In this paper, inspired by [25], we adopted Bayesian Network (BN) to represent and capture the semantic relationships among AUs, as well as the correlations of the AU intensities, in order to more accurately and robustly measure the intensity of spontaneous facial actions.

4.1.2. BN structure learning

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. In this work, we employ 12 hidden nodes representing 12 AUs of DISFA database [9], (i.e., AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25, AU26), each of which has six discrete states indicating the intensity of the AU. The structure of a BN captures the dependency relationships among variables, and is crucial for accurately modeling the joint probability distribution. Hence, we employ BN structure learning algorithm to learn the BN structure from both training data and prior knowledge.

The goal of BN structure learning is to find a structure B that maximizes a score function such as Bayesian Information Criterion (BIC) [26] score function:

$$\max_B S_B(\theta) = \max_B \left( L_B(\theta) - d \cdot \frac{1}{2} \log K \right)$$

(1)

where $L_B(\theta)$ is the log-likelihood function of parameter $\theta$ with respect to data $D$ and structure $B$, $d$ is the number of free parameters in $B$ and $K$ is the number samples in training data. $L_B(\theta)$ can be expressed as

$$L_B(\theta) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \prod_{k=1}^{r_i} \theta_{ijk}^{B}$$

(2)

where $\theta_{ijk}^{B}$ represents the parameters for network $B$, $n$ is the number of variables (nodes), $m_i$ is the number of parent configuration of $i$th node and $r_i$ is the number of possible states of $i$th node. Note that the value of $m_i$ depends on the structure $B$. More specifically, it depends on the parent set of each node. The number of free parameters in $B$ is computed as $d = \sum_{i=1}^{n} m_i (r_i - 1)$.

Based on the score function, the BN structure learning can be defined as an optimization problem. Campos et al. [27,28] recently developed an exact BN structure learning algorithm from data and expert's knowledge based on BIC score function. They [27,28] found some properties that strongly reduce the time and memory costs, and developed a branch and bound algorithm that integrates structural constraints with data in a way to guarantee global optimality. In this work, we employ the method presented in [28] to learn our BN structure from both training data and prior knowledge. The prior knowledge is derived from both data analysis and study of previous works.

Table 1 calculates the Pearson Correlation Coefficient (PCC) of the AU intensities on DISFA database. From Table 1 we can see that there are strong correlation between some variables. For instance the PCC between the intensities of AU2 and AU1 is 0.732, which means that there is strong correlation relationship between the intensities of these two AUs, e.g., high/low intensity of AU2

![Fig. 4. AU combinations that show meaningful expression. (a) AU6 + AU12 + AU25 to represent happiness. (b) AU4 + AU15 + AU17 to represent sadness. (c) AU9 + AU17 to represent disgust. (Adapted from [9]).](image-url)
indicates a high probability of high/low intensity of AU1. The strong correlation relationship between two variables can be modeled by a link in BN. Thus, if the PCC between the intensities of two AUs is higher than \( t \) (\( t \) is a threshold and we set as 0.5 in this work), we add a link (a constraint) between the two AUs in the BN to model such correlation relationship. The direction of the link is set based on the characters of the AUs we are going to recognize, as well as the analysis of previous studies [38,25]. This way, we manually constructed three constraints, i.e., a link from AU2 to AU1, a link from AU12 to AU6 and a link AU12 to AU25, which will help reduce the search space. The manually constructed constraints are reasonable for understanding spontaneous facial expressions. For instance, in spontaneous expression, AU6, AU12 and AU25 often co-occur to represent happiness, and the intensities of the three AUs interact with each other to represent the intensity of the happiness. Hence, there are links among these three variables in the BN to model such correlation relationships.

Given the complete training data and the structural constraints, we employ the learning algorithm presented in [28] to learn our BN structure. The learned structure is shown in Fig. 5. From Fig. 5 we can see that the learned structures such as AU9 to AU6, AU2 to AU5, and AU17 to AU15 are all reasonable to reflect facial expressions. For example, pain expression often involves AU6 and AU9 which means that there is high co-occurrence relationship between these two AUs. Similarly, AU2 and AU5 often occur together in surprise and fear expressions, and sadness expression usually involves AU15 and AU17. In summary, the learnt structure is capable of reflecting the pattern of the AU intensity correlation relationships in the training set.

### 4.2. Dynamic dependencies learning

The above BN structure can only capture the static dependencies among AU intensities. In this section, we extend it to a dynamic Bayesian network by adding dynamic links. In general, a DBN is made up of interconnected time slices of static BNs based on two assumptions: first, the system is the first-order Markovian, i.e., \( P(x^{t+1} | x_t) = P(x^{t+1} | x_t^t) \). Another assumption is that the transition probability \( P(x^{t+1} | x_t) \) is the same for all the \( t \). Therefore, a DBN can be defined by a pair of BNs \((B_1, B_2)\): (1) the static network \( B_1 \) is the same as what we learned above, which captures the static relationships among AU intensities; (2) the transition network \( B_2 \) specifies the transition probability \( P(x^{t+1} | x_t) \) for all \( t \) in a finite time slice sequence, i.e., the dynamic dependencies among AU intensities.

Then, given a DBN model, the joint probability over the sequence \( x^1, \ldots, x^T \) can be computed as follows:

\[
P(x^1, \ldots, x^T) = P_\text{B}_1(x^1) \prod_{t=1}^T P_\text{B}_2(x^{t+1} | x^t)
\]  

where \( P_\text{B}_1(x^1) \) captures the joint probability of variables in the static BN \( B_1 \), and \( P_\text{B}_2(x^{t+1} | x^t) \) captures the transition probability and can be decomposed based on the transition network:

\[
P_\text{B}_2(x^{t+1} | x^t) = \prod_{i=1}^n P_{\text{B}_2}(x_{t+1}^i | \text{pa}(x_{t+1}^i))
\]  

where \( \text{pa}(x_{t+1}^i) \) represents the parent configuration of variable \( x_{t+1}^i \) in the transition network \( B_2 \), and \( n \) indicates the number of random variables in \( X^t \).

As discussed above, a DBN can be defined by a pair of BNs \((B_1, B_2)\). The static network \( B_1 \) has been learned in the above sections. Here, we focus on learning the transition network \( B_2 \). Given the complete training data, the learning process of \( B_2 \) is the same as the static BN structure learning. We just need to reformat the two successive frames as on data instance to train the transition network \( B_2 \).

But compared to the static network, there are more coherent structural constraints on the transition network. The transition network consists of two types of links: inter-slice links and intra-slice links. The inter-slice links are the dynamic links connecting the temporal variables of two successive time slices. In contrast, the intra-slice links connect the variables within a single time slice, which are same as the static network structure. The coherent structural constraints on the transition network are as follows: First, the variables in the first time slice do not have parents. Second, the inter-slice links can only have one direction, in other words from the previous time slice to the current time slice. Finally, based on the stationary assumption, the intra-slice links...
for every time slice should be the same. Furthermore, to reduce the searching space and increase the generalization ability, we impose an additional constraint such that each node $X_{it}^{j+1}$ has at most two parents from the previous time slice. Besides, we also add some experiential constraints based on analysis of previous studies. Got the complete training data and the above structural constraints, we then apply the learning algorithm in [28] to identify the transition network structure. Here, we fixed the intra-slice links as the previously learned static network.

The learned transition network is shown in Fig. 6(a). For presentation clarity, we use the self-arrows at each AU node to indicate the temporal dependencies of a single AU from the previous time slice to the current time slice. From Fig. 6(a) we can see that the learned transition network felicitously reflects the dynamic dependencies among AU intensities. For example, the dynamic link from $AU_{6}^{t-1}$ to $AU_{6}^{t}$ means that the eyebrows intend to lower by activating AU4 as the intensity of cheek raiser (AU6) increases. And the dynamic link from $AU_{26}^{t-1}$ to $AU_{6}^{t}$ means that before the cheek is raised by activating AU6 (cheek raiser), it is most likely the lip corners are already pulled upward by activating AU12 (lip corner puller).

Finally, Fig. 6(b) shows the completed DBN model for AU intensity recognition. The unshaded nodes are the hidden nodes, whose states represent AU intensities and need to be inferred from the DBN, given the image observations. The image observations, which are obtained from the computer vision techniques discussed in Section 3, are represented using the shaded nodes. The links from the hidden nodes to the observation nodes capture the likelihood of the image observation given the corresponding AU intensity, e.g., $P(\text{MAU}_1|\text{AU}_1)$.

### 4.3. DBN parameter learning

Given the DBN structure, now we focus on learning the parameters from training data to infer the hidden nodes. During the DBN parameter learning, we treat the DBN as an expanded BN consisting of two-slice static BNs connected through the temporal variables. Learning the parameters in a DBN is to find the most probable values $\hat{\theta}$ for $\theta$ that can best explain the training data. Let $\theta_{ijk}$ indicate a probability parameter:

$$\theta_{ijk} = P(X_{it}^{j}|pa(X_i))$$

where $i$ ranges over all the variables (nodes in the BN), $j$ ranges over all the possible parent instantiations for variable $X_i$, and $k$ ranges over all the instantiations for $X_i$ itself (intensity levels of AUs). Therefore, $X_{it}^{j}$ represents the $k$th state of variable $X_i$, and $\text{pa}(X_i)$ 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 that the training data are independent, based on the conditional independence assumption in DBN, we have the log likelihood function in Eq. (6), 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:

$$L_D(\theta) = \log \prod_{i=1}^{n} \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{ijk}$$

Since we have a complete training data, i.e., for each frame we have the intensity labels for all 12 AUs, the learning process can be described as a constrained optimization problem as follows:

$$\arg \max L_D(\theta), \text{ s.t. } g_{ij}(\theta) = \sum_{k=1}^{K} \theta_{ijk} - 1 = 0$$

where $g_{ij}$ imposes that distributions defined for each variable given a parent configuration sums to one over all variable states. This problem has its global optimum solution at $\theta_{ijk} = n_{ijk}/\sum_{k=1}^{K} n_{ijk}$.

To evaluate the quantity of training data needed for learning the proposed model, we perform a sensitivity study of model learning on different amount of training data. For this purpose, the Kullback–Leibler (KL) divergences of the parameters are computed versus the number of training samples as shown in Fig. 7. From Fig. 7 we can see that the learning process of the proposed model requires a total of about 50 000 training samples. Since we have more than 50 000 training samples available in the database, the training data for the model learning are sufficient.

### 4.4. DBN inference

Given the complete DBN model and the AU image observations, we can estimate the AU intensities by maximizing the posterior probability of the hidden nodes. Let $AU_{1:N}^{t}$ represent the nodes for $N$ target AUs at time $t$. Given the available evidence until time $t$: $\text{PAU}_{1:N}^{t-1}$, the posterior probability $P(AU_{1:N}^{t}|\text{PAU}_{1:N}^{t-1})$ can be factorized and computed via the proposed model by performing the DBN inference.
it has multiple targets [32]. In other words, ICC measures the correlation or conformity for a data set when one judges or measurement devices. The ICC in is defined as

\[
\text{ICC} = \frac{(BMS - EMS)}{BMS + (k - 1) \times EMS}
\]

where BMS is the between-targets mean squares and EMS is the residual mean squares defined by Analysis Of Variance (ANOVA). That is, the ICC indicates the proportion of total variance due to differences between targets. See [32] for additional details.

5. Experimental results

In our experiments we utilized the DISFA database for evaluating the performance of automatic measurement of the intensity of spontaneous action units. First we introduce the contents of DISFA and then the results of the proposed system for measuring the intensity of 12 AUs of this database are reported.

5.1. DISFA database description

Denver Intensity of Spontaneous Facial Action (DISFA) database [9,33] contains the 4-min videos of spontaneous facial expressions for 27 adults participants. The facial images of DISFA were video recorded by a high resolution camera (i.e. 1024 × 768 pixel) at a frame rate of 20 fps. In DISFA, the intensity of 12 AUs (i.e. AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU20, AU25, AU26 ) has been coded by a FACS coder and the six levels of AU intensities. The 6-level intensity in DISFA can represent the dynamics of each AU by the ordinal scale intensity, 0 (absence of an AU) to 5 (maximum intensity) [3]. The database also provides a set of 66 landmark points that represent the coordinates of important components of human’s face, such as corner of the eyes and boundary of the lips. [9]. In this study, we utilized entire set of video frame of all participants to measure the intensity of 12 aforementioned AUs. Table 2 (adopted from [9]) lists the total number of frames for each AU intensity level on DISFA database.

5.2. Recognition and reliability measures

To evaluate the proposed automated AU intensity measurement we exploited both recognition rate and Intra-Class Correlation (ICC) value. ICC is a statistical index that ranges from 0 to 1 and is a measure of correlation or conformity for a data set when it has multiple targets [32]. In other words, ICC measures the reliability studies in which n targets of data are rated by k judges (i.e., in this paper k=2 and n=6). ICC is similar to Pearson correlation and is preferred when computing consistency between judges or measurement devices. The ICC in is defined as

\[
\text{ICC} = \frac{(BMS - EMS)}{BMS + (k - 1) \times EMS}
\]

where BMS is the between-targets mean squares and EMS is the

Table 2

<table>
<thead>
<tr>
<th>AU no.</th>
<th># Frames for each AU intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>122 036 227 2749 2809 1393 555</td>
</tr>
<tr>
<td>1</td>
<td>123 450 1720 934 3505 836 369</td>
</tr>
<tr>
<td>2</td>
<td>106 220 4661 7636 6586 4128 1383</td>
</tr>
<tr>
<td>3</td>
<td>128 085 1579 719 293 104 34</td>
</tr>
<tr>
<td>4</td>
<td>111 330 9157 5986 3599 601 141</td>
</tr>
<tr>
<td>5</td>
<td>123 682 1659 2035 3045 316 77</td>
</tr>
<tr>
<td>6</td>
<td>100 020 13943 6869 7233 2577 172</td>
</tr>
<tr>
<td>7</td>
<td>122 952 5180 1618 1017 47 0</td>
</tr>
<tr>
<td>8</td>
<td>117 884 6342 4184 2281 112 11</td>
</tr>
<tr>
<td>9</td>
<td>126 282 1591 1608 1305 28 0</td>
</tr>
<tr>
<td>10</td>
<td>84 762 9805 13 935 15 693 5580 1039</td>
</tr>
<tr>
<td>11</td>
<td>105 838 13 443 7473 3529 314 217</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Feature type</th>
<th>SVM</th>
<th>BN</th>
<th>DBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG feature</td>
<td>66.64</td>
<td>68.95</td>
<td>70.16</td>
</tr>
<tr>
<td>Gabor feature</td>
<td>76.57</td>
<td>77.71</td>
<td>78.34</td>
</tr>
</tbody>
</table>

Fig. 7. The KL divergence of the proposed DBN model versus the training data size. We use different number of subject’s data to train the model.

To demonstrate the correlation relationships among AUs intensities, we list the correlation matrix between AU2 and AU1 on DISFA database in Table 4. From Table 4 we can see that the AU
intensity dependency relationship between AU2 and AU1 is strong, i.e., high/low intensity of AU2 indicates a high probability of high/low intensity of AU1. For example, when the intensity of AU2 is at level “0”, the probability of AU1 = 0 is 0.9710, and when the intensity of AU2 is at level “5”, the probability of AU1 = 5 is 0.8969. By modeling such AU intensity dependency relationship between AU2 and AU1 in the DBN model, for the HOG features, the ICC of AU1 is increased from 69.31 percent (for the AdaBoost classifier) to 72.77 percent (for the DBN model), and that of AU2 is increased from 56.01 percent to 63.48 percent, and for the Gabor feature, the ICC of these two AUs is increased from 79.68 percent to 82.90 percent, and from 82.68 percent to 87.33 percent respectively.

Similarly, we list the correlation matrix between AU9 and AU5 in Table 5. From Table 5 we can see that, when AU9 occurs, the probability of AU5 being present is low, which means that there is exclusive relationship between these two AUs. The proposed model can improve the observation accuracy by modeling such relationships, especially for image observations with low accuracy. For example, for HOG feature observation, AU5 is not well recognized by the SVM, and the proposed model increases the probability of AU5 being present is low, which means that there is exclusive relationship between these two AUs. The proposed model can improve the observation accuracy by modeling such

Finally, Table 6 lists the recognition results (ICC and accuracy) for each intensity level by using the proposed DBN model on DISFA database. Accuracy was defined as an average of correctly detected intensity levels for each AU. From Table 6 we can see that the average recognition accuracy for AU intensity at levels “1” and “5” is lower than that for AU intensity at other levels. The low recognition accuracy for AU intensity level “1” is mainly because that the low intensity expression produces very subtle facial changes, and is really hard to detect. However, high intensity level represents exaggerated expression which intuitively should be easily recognized. The low recognition accuracy for AU intensity “5” is most likely because of lack of training samples. For example, Table 2 (Adopted from [9]) lists the total number of frames for each AU intensity level on DISFA database, and from Table 2 we can see that the number of frames for intensity level “5” is extremely smaller than that for other levels.

Table 7 lists the comparison of our work with previous state of the art method for measuring the intensity of AUs on DISFA database, in terms of both ICC and accuracy. Note that AU intensity recognition is a relatively recent problem within the field, and there are few works measuring the intensity of AUs. Mavadati et al. presented the DISFA database in [9], and provided a baseline method to measuring AU intensity. From Table 7 we can see that the proposed method achieves slight better results compared to

![Fig. 8. The comparison of AU intensity recognition results by using the SVM classifier alone, the BN, and the DBN on DISFA database. (a) Recognition results for HOG feature. (b) Recognition results for Gabor feature.](image)
Table 6
The recognition results for each intensity level by using the proposed DBN model on DISFA database.

<table>
<thead>
<tr>
<th>Feature type</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU5</th>
<th>AU6</th>
<th>AU9</th>
<th>AU12</th>
<th>AU15</th>
<th>AU20</th>
<th>AU25</th>
<th>AU26</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) HOG features</td>
<td>72.77</td>
<td>63.48</td>
<td>86.67</td>
<td>57.05</td>
<td>70.03</td>
<td>75.27</td>
<td>86.53</td>
<td>62.29</td>
<td>57.73</td>
<td>54.73</td>
<td>85.07</td>
<td>70.31</td>
</tr>
<tr>
<td>ICC (%)</td>
<td>82.19</td>
<td>82.81</td>
<td>79.77</td>
<td>92.95</td>
<td>75.49</td>
<td>90.26</td>
<td>77.57</td>
<td>86.09</td>
<td>76.24</td>
<td>90.92</td>
<td>68.85</td>
<td>69.88</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>98.58</td>
<td>98.81</td>
<td>98.07</td>
<td>99.64</td>
<td>95.89</td>
<td>99.26</td>
<td>95.44</td>
<td>98.04</td>
<td>97.26</td>
<td>99.44</td>
<td>93.26</td>
<td>94.22</td>
</tr>
<tr>
<td>6-Level</td>
<td>'0'</td>
<td>4.27</td>
<td>4.25</td>
<td>11.00</td>
<td>11.79</td>
<td>17.95</td>
<td>9.90</td>
<td>31.46</td>
<td>12.46</td>
<td>12.67</td>
<td>8.38</td>
<td>18.02</td>
</tr>
<tr>
<td>'1'</td>
<td>18.07</td>
<td>15.35</td>
<td>40.53</td>
<td>24.35</td>
<td>41.44</td>
<td>24.61</td>
<td>45.80</td>
<td>20.84</td>
<td>23.17</td>
<td>29.85</td>
<td>50.30</td>
<td>29.77</td>
</tr>
<tr>
<td>'2'</td>
<td>47.24</td>
<td>29.93</td>
<td>49.55</td>
<td>40.82</td>
<td>44.38</td>
<td>71.55</td>
<td>72.75</td>
<td>60.16</td>
<td>66.09</td>
<td>26.20</td>
<td>72.64</td>
<td>40.59</td>
</tr>
<tr>
<td>'3'</td>
<td>46.68</td>
<td>29.16</td>
<td>61.18</td>
<td>20.74</td>
<td>26.24</td>
<td>66.77</td>
<td>61.33</td>
<td>17.65</td>
<td>6.25</td>
<td>5.56</td>
<td>57.61</td>
<td>36.89</td>
</tr>
<tr>
<td>'4'</td>
<td>33.18</td>
<td>14.58</td>
<td>61.28</td>
<td>38.46</td>
<td>0</td>
<td>55.41</td>
<td>0</td>
<td>14.29</td>
<td>-</td>
<td>60.53</td>
<td>61.41</td>
<td>28.18</td>
</tr>
<tr>
<td>'5'</td>
<td>82.90</td>
<td>87.33</td>
<td>88.14</td>
<td>80.89</td>
<td>80.99</td>
<td>80.59</td>
<td>84.99</td>
<td>72.20</td>
<td>74.14</td>
<td>55.93</td>
<td>94.75</td>
<td>80.70</td>
</tr>
</tbody>
</table>

Table 7
Comparison with state of the art method.

<table>
<thead>
<tr>
<th>Works</th>
<th>HOG features</th>
<th>Gabor features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Work [9] 2013</td>
<td>0.70</td>
<td>79.14</td>
</tr>
<tr>
<td>Our work</td>
<td>0.7016</td>
<td>81.08</td>
</tr>
</tbody>
</table>

6. Conclusions and future work

Due to the richness, ambiguity, and dynamic nature of facial actions, individually and statically recognizing each AU intensity is not always accurate and reliable for spontaneous facial expressions. Hence, improving the recognition system's efficiency not only requires improving the observation extraction accuracy, but more importantly, requires exploiting the spatiotemporal interactions among facial actions, since it is these coherent, coordinated, and synchronized interactions that produce a meaningful facial display. In this work, we presented a DBN based probabilistic framework to model the semantic and dynamic relationships among multilevel AU intensities. Extracted the image observations, AU intensities recognition is accomplished through probabilistic inference by systematically integrating the image observations with the proposed DBN model. In this way, the erroneous AU intensity observations can be compensated by the model's built-in spatial and temporal relationships, thus improving the efficiency of AU intensity measurement system. For instance, experimental results on DISFA database show that the overall ICC value for HOG features increased from 66.64% to 70.17% and for Gabor feature from 76.57% to 78.31%. As in spontaneous facial expression several factors, such as head pose variation, might happen. In order to deal with such challenges to measure the intensity of spontaneous action units, as a future work, one might introduce another hidden layer node to model to accommodate those challenges.

Conflict of interest

None declared.

Acknowledgements

This work was mainly accomplished when the first author visited Rensselaer Polytechnic Institute (RPI) as a visiting student, and was partially supported by National Natural Science Foundation of China Funded Project 61402129, Helongjiang Postdoctoral Science Foundation Funded Project LBH-Z14090, awards IIP-1111568 and BCS-1052781 from the National Science Foundation to the University of Denver.

References

Yongqiang Li received the B.S., M.S. and Ph.D. degrees in instrument science and technology from Harbin Institute of Technology, Harbin, China, in 2007, 2009 and 2014, respectively. He is currently an Assistant Professor at Harbin Institute of Technology. He worked as a visiting student at Rensselaer Polytechnic Institute, Troy, USA, from September 2010–September 2012. His areas of research include computer vision, pattern recognition, and human-computer interaction.

S. Mohammad Mavadati received the B.Sc. degree in electronics engineering from the Shahrood University of Technology, Iran, in September 2007, and the M.Sc. degree in telecommunication engineering from Yazd University, Iran, in March 2010. He is currently working toward the Ph.D. degree and is a Graduate Research Assistant in the Department of Electrical, Computer, and Instrument Science and Technology at Harbin Institute of Technology, Harbin, China. His areas of research include signal processing, system integration and pattern classification, and computer vision. He is a student member of both the IEEE and the IEEE Signal Processing Society.

Mohammad H. Mahoor received the B.S. degree in electronics from the Abadan Institute of Technology, Iran, in 1996, the M.S. degree in biomedical engineering from the Sharif University of Technology, Iran, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of Miami, Florida, in 2007. He joined the University of Denver (DU) as an Assistant Professor of computer engineering in September 2008. He has authored or coauthored more than 60 refereed research publications. He is the director of image processing and computer vision laboratory at DU. His research interests include affective computing and developing automated systems for facial expression recognition. He is a member of the IEEE.

Yongping Zhao received the Ph.D. degree in electrical engineering from Harbin Institute of Technology, Harbin, China. He is currently a Professor with the Department of Instrument Science and Technology at Harbin Institute of Technology, Harbin, China. His areas of research include signal processing, system integration and pattern recognition.

Qiang Ji received his Ph.D. degree in electrical engineering from the University of Washington. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He recently served as a Program Director at the National Science Foundation (NSF), where he managed NSF’s computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robert Packate Institute at Rensselaer Polytechnic University, the Department of Computer Science at University of Nevada at Reno, and the US Air Force Research Laboratory. Prof. Ji currently serves as the Director of the Intelligent Systems Laboratory (ISL) at RPI. Prof. Ji’s research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published over 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including NSF, NIH, DARPA, ONR, ARO, and AFOSR as well as major corporations including Honda and Boeing. Prof. Ji is an Editor on several related IEEE and international journals and has served as a General Chair, Program Chair, Technical Area Chair, and Program Committee Member in numerous international conferences/workshops. Prof. Ji is a Fellow of IEEE and IAPR.