Pattern Recognition 61 (2017) 78–91

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Expression-assisted facial action unit recognition under incomplete AU annotation

Shangfei Wang [a,*], Quan Gan [a], Qiang Ji [b]

[a] Key Lab of Computing and Communication Software of Anhui Province, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, PR China

[b] Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form
9 July 2016
Accepted 18 July 2016
Available online 21 July 2016

Keywords:
AU recognition
Incomplete annotation
Bayesian network
Expression

ABSTRACT

Facial action unit (AU) recognition is an important task for facial expression analysis. Traditional AU recognition methods typically include a supervised training, where the AU annotated training images are needed. AU annotation is a time consuming, expensive, and error prone process. While AU is hard to annotate, facial expression is relatively easy to label. To take advantage of this, we introduce a new learning method that trains an AU classifier using images with incomplete AU annotation but with complete expression labels. The goal is to use expression labels as hidden knowledge to complement the missing AU labels. Towards this goal, we propose to construct a Bayesian network (BN) to capture the relationships among facial expressions and AUs. Structural expectation maximization (SEM) is used to learn the structure and parameters of the BN when the AU labels are missing. Given the learned BNs and measurements of AUs and expression, we can then perform AU recognition within the BN through a probabilistic inference. Experimental results on the CK+, ISL and BP4D-Spontaneous databases demonstrate the effectiveness of our method for both AU classification and AU intensity estimation.

1. Introduction

Facial expression recognition has attracted increasing attention due to its wide applications in human–computer interaction [1]. There are two kinds of descriptors of expressions: expression category and AUs [2]. The former describes facial behavior globally, and the latter represents facial muscle actions locally. To recognize AUs and expressions, a large number of annotated training images are required. In general, AU annotation is more expensive and harder than expression annotation, since expression is global and easier to recognize, while AUs are local and subtle, and harder to recognize. Furthermore, the number of AUs for an image is usually larger than that of expressions. Therefore, the AUs should be labeled by qualified facial action coding system (FACS) experts. Current research [3–6] reveals that some AUs are obvious and easy to be annotated, while others are subtle and hard to annotate. This phenomenon not only increases the difficulty of AU annotation, but also makes the AU labels error prone. Compared with AU annotation, expressions are much easier to annotate, and can be labeled with great accuracy.

In this work, we try to design an AU recognition method with the assistance of expression labels under the incomplete AU labeling. Specifically, during training, instead of trying to label every AU in each image as being done by the existing AU recognition methods, we only label AUs that can be labeled confidently and leave those difficult and subtle AUs unlabeled. In addition, we provide expression label for each image. Using such annotated images, we then train an AU recognition algorithm by leveraging on the relationships among AUs and the knowledge of the expressions. To take advantage of the available expression labels during training, we propose to construct a BN to systematically capture the dependencies among AUs and expressions. The nodes of the BN represent the AUs and expressions. The links and their parameters capture the probabilistic relations among AUs and expressions. Since some AU labels are missing for some training images, structural expectation maximization (SEM) is adopted to learn the structure and parameters of the BN. Given the learned BN, we can infer the AUs by combining the AU-expression relationships encoded in the BN and the AU measurements. The experimental results on the CK+ database show that, with complete annotation, our method outperforms the state of the art model-based and image-driven AU classification methods; with incomplete annotation, our method performs much better than state of the art AU classification methods. The experimental results on the ISL database demonstrate the cross-database generalization

* Corresponding author.
E-mail addresses: sfwang@ustc.edu.cn (S. Wang), gqquan@mail.ustc.edu.cn (Q. Gan), qji@ecse.rpi.edu (Q. Ji).

http://dx.doi.org/10.1016/j.patcog.2016.07.028
0031-3203/

ability of our method for AU classification. Furthermore, the experimental results on the BP4D-Spontaneous databases demonstrate that for AU intensity estimation, our method outperforms current model-based and image-driven AU intensity estimation methods under both complete and incomplete annotation.

2. Related work

Usually, several AUs can be present at the same image or image sequence. Thus, AU recognition can be formulated as a multi-label classification problem. Due to the large number of possible label sets, multi-label classification is rather challenging. Successfully exploiting the dependencies inherent in multiple labels is the key to facilitate the learning process. Accounting for dependencies among AUs, present AU recognition research can be divided into three groups.

The first group recognizes each AU individually and directly from images or sequences [7,8]. They are referred to as image-based AU recognition methods. Valstar and Pantic [7] proposed an automatic method to detect 22 AUs. They first detected and tracked 20 facial points, and then used a combination of GentleBoost, support vector machines (SVMs), and hidden Markov models as a classifier. van der Maaten and Hendriks [8] adopted AAM features and linear chain conditional random field to detect the presence of AUs. These works treat recognition of each AU individually as one-vs.-all scheme, ignoring the dependencies among AUs. However, multiple AUs can appear together and thus there exist dependencies among them. The AU relationships may help AU recognition.

The second group recognizes fixed AU combinations. One approach regards the AU combination as a new AU. For example, Littlewort et al. [9] analyzed the AU combinations of 1+2, 2+4, 1+4 and 1+2+4 using a linear SVM with Gabor features. Lucey et al. [10] used SVM and nearest neighbor to detect a few combinations of AUs (i.e. 1, 1+2, 4, 5) with active appearance model (AAM) features. The other approach integrates AU relations existing in AU labels into AU classifiers. Zhang and Mahoor [11] first proposed a hierarchical model to group multiple AUs into several fixed groups based on AU co-occurrences existing in AU labels and facial regions. Then, each AU recognition is regarded as a task, and AUs in the same groups share the same kernel. A multi-task multiple kernel learning is used to learn AU classifiers simultaneously. Zhao et al. [12] selected a sparse subset of facial patches and learned multiple AU classifiers simultaneously under the constraints of group sparsity and local AU relations (i.e. positive correlation and negative competition). Although the local dependencies among AUs have been exploited in these works, the combinations are manually determined and fixed. Thus, it is only feasible for a few combinations and is hard to detect thousands of possible combinations.

The third group explicitly exploits the co-existent and mutually exclusive relations among AUs from target labels. They are referred to as model-based AU recognition methods. Tong et al. [13,14] used Gabor features and SVM to recognize each AU first, then they model the relations among AU labels by dynamic Bayesian network (DBN). Eleftheriadis et al. [15] proposed a multi-conditional latent variable model to combine global label dependencies into latent space and classifier learning. The image features are projected onto the latent space, which is regularized by constraints, encoding local and global co-occurrence dependencies among AU labels. Then, multiple AU classifiers are learned simultaneously on the manifold. Both work assumes complete AU labeling and only involves AUs without using expressions. Wang et al. [16] proposed a hierarchical model to integrate the low-level image measurements with the high-level AU semantical relationships for AU recognition. A restricted Boltzmann machine (RBM) is used to capture higher-order AU interactions, and a 3-way RBM is further developed to capture related factors such as the facial expressions to achieve better characterization of the AU relations. Although their model can capture high order AU relationships, it cannot effectively handle missing labels. Therefore, all the model-based AU recognition methods require completely annotated images.

Due to the difficulties of collecting data with AU intensity values and the limited available database, current AU analyses mainly focus on AU occurrence detection, and few work measures the intensity of AUs. Furthermore, among the very few existing AU intensity estimation work, most work, such as [17–23], measures the intensity of each AU independently. They do not make use of the intensity dependencies that are crucial for analyzing AUs. In this paper, we refer to these methods as image-driven intensity estimation methods. It is only recently that two works have considered AU relations for AU intensity estimation. Li et al. [24] proposed using DBN to model AU relationships for measuring their intensities. In order to estimate the intensity of AUs present in a region of the upper face, Sandbach et al. [25] adopted Markov random field structures to model AU combination priors. Similar to Tong et al.'s [13,14] work, both works assume complete AU intensity labeling and only involve AUs without using expressions. They are referred to as model-based AU intensity estimation methods. Therefore, all the model-based AU intensity estimation methods [24,25] require completely AU-annotated images without using expressions.

To the best of our knowledge, there is little reported work that recognizes AUs or estimates AU intensities with the assistance of expressions [16,26], although there exist a few works considering the relations among expressions and AUs to help expression recognition or to jointly recognize AUs and expressions [27]. For example, Pantic and Rothkrantz [28] summarized the production rules of expressions from AUs using the AUs-coded descriptions of the six basic emotional expressions given by Ekman and Friesen [2]. Velusamy et al. regarded AU to expression mapping as a problem of approximate string matching, and they adopted a learned statistical relationship among AUs and expressions to build template strings of AUs for six basic expressions [5]. Zhang and Ji proposed to use DBNs to model the probabilistic relations of facial expressions to the complex combination of facial AUs and temporal behaviors of facial expressions [6]. Li et al. [27] introduced a dynamic model to capture the relationships among AUs, expressions and facial feature points. The model was used to perform AU and expression recognition as well as facial feature tracking.

Current AU recognition and AU intensity estimation methods require complete AU label assignments. However, AU analyses with incomplete AU label assignments are frequently encountered in realistic scenarios, due to the large number of AUs and difficulty in manual AU annotation. Till now, little research has addressed the challenge of AU analyses with incomplete AU labels. While AUs are hard to annotate, facial expression is relatively easy to label. The available expression labels during training and the dependencies among expression and AUs may be useful for AU recognition and AU intensity estimation. Thus, in the paper, we construct AU classifiers with the ability of learning and inferring from incomplete AU annotations with the help of ground truth of expression knowledge that is available during training only.

Compared with related work, the main contribution of this work lies in the introduction of a probabilistic framework to use the ground truth of expression labels available during training to help train an improved AU classifier under incomplete AU annotation. In addition, we formulate AU detection as a multi-label classification problem.

3. Methods

Assuming the training data is $D = \left\{ X_j, (\lambda_{1j}, \ldots, \lambda_{nj}, \lambda_{n+1j}) \right\}_{j=1}^{m}$, in which m is the number of training samples, n is the number of AU labels. For the jth training image, we denote X_j as the image feature vector, and $\lambda_{1j}, \ldots, \lambda_{nj}$ as the n AU labels, and λ_{n+1j} as the expression label respectively. All the other data are complete, except for the AU label data, which may be missing due to the hard annotation. The ground truth of expression label λ_{n+1j}, while available during training, is unavailable during testing. The goal of this work is to construct an AU classifier or an AU estimator with the abilities of learning and inferring from incomplete AU data with the help of expression knowledge that is available during training. As shown in Fig. 1, our approach consists of two modules: AU and expression measurement extraction, and AUs and expression' relations modeling by BN. The training phase of our approach includes training the traditional image-based methods for AU measurement extraction and training the BN to capture the semantic relationships among AUs and expressions. Given the measurements, we infer the final labels of samples through the most probable explanation (MPE) inference with the BN model. The details are provided as follows.

3.1. Measurement extraction

The measurements $\hat{\lambda}$ are the preliminary estimations of the AU and expression labels using an existing image-driven recognition method based on training data. In this work, the face registration is performed using the detected feature points firstly, and then the movements of the feature point between the neutral and apex images are used as the image features. After that, for AU recognition, a SVM is used as the classifiers to obtain the initial AU measurements and expression measurements. For AU intensity estimation, multi-class SVM and support vector regression (SVR) have been used as the predictors to extract expression measurements and AU measurements respectively.

3.2. Expression-dependent AU recognition

In order to model the semantic relationships among AUs and expressions, a BN model is utilized in this work. As a probabilistic graphical model, BN can effectively capture the dependencies among variables in data. In our work, each node of the BN is an AU or expression label, and the links and their conditional probabilities capture the probabilistic dependencies among AUs and expressions, as shown in Fig. 2.

3.2.1. BN structure and parameters learning for incomplete Aus' label

A BN is a directed acyclic graph (DAG) $G = (\Lambda, E)$, where $\Lambda = \left\{ \lambda_i \right\}_{i=1}^{n+1}$ represents a collection of $n+1$ nodes and E denotes a collection of arcs.

Given the dataset of multiple target labels $TD = \{\lambda_{ij}\}$, where $i = 1, 2, \ldots, n, n+1$ is an index to the number of nodes, and $j = 1, 2, \ldots, m$ is index to the number of samples. The structure and parameter learning is to find a structure G that maximizes a score function. In this work, we employ the Bayesian information criterion (BIC) score function which is defined as Eq. (5), and the BN structural learning algorithm proposed by de Campos and Ji [29] is employed.

By exploiting the decomposition property of the BIC score function, this method allows learning an optimal BN structure efficiently and it guarantees to find the global optimum structure, independent of the initial structure. Furthermore, the algorithm provides an anytime valid solution, i.e. the algorithm can be stopped at any-time with a best current solution found so far and an upper bound to the global optimum. Representing state of the art method in BN structure learning, this method allows automatically capturing the relationships among expressions. Details of this algorithm can be found in [29]. Examples of the trained BN structure are shown in Figs. 2 and 4.

After the BN structure is constructed, parameters can be learned from the training data. Learning the parameters in a BN means finding the most probable values $\hat{\theta}$ for θ that can best explain the training data. Here, let λ_i denotes a variable of BN, and θ_{ilk} denotes a probability parameter for BN, then,

$$\theta_{ilk} = P\left(\lambda_i^k | pa^l\left(\lambda_i\right)\right) \tag{1}$$

where $i \in \{1, \ldots, n+1\}$, $l \in \{1, \ldots, r_i\}$ and $k \in \{1, \ldots, s_i\}$. Here n denotes the number of variables (nodes in the BN); r_i represents the number of the possible parent instantiations for variable λ_i, i.e. the number of the possible instantiations of $pa(\lambda_i)$; s_i indicates the number of the state instantiations for λ_i. Hence, λ_i^k denotes the kth state of variable λ_i.

Based on the Markov condition, any node in a BN is conditionally independent of its non-descendants, given its parents. The joint probability distribution represented by BN can be denoted as: $P(\lambda) = P(\lambda_1, \ldots, \lambda_{n+1}) = \prod_i P(\lambda_i | pa(\lambda_i))$. In this work, the "fitness" of parameters θ and training data D is quantified by the log likelihood function $\log(P(D|\theta))$, denoted as $L_D(\theta)$. Assuming the training data are independent, based on the conditional independence assumptions in BN, the log likelihood function is shown in the following equation:

$$L_D(\theta) = \log\left(\prod_{i=1}^{n+1} \prod_{l=1}^{r_i} \prod_{k=1}^{s_i} \theta_{ilk}^{w_{ilk}} \right) \tag{2}$$

where w_{ilk} indicates the number of elements in D containing both λ_i^k and $pa^l(\lambda_i)$.

For fully labeled training data, maximum likelihood estimation (MLE) method can be described as a constrained optimization problem, which is shown in the following equation:

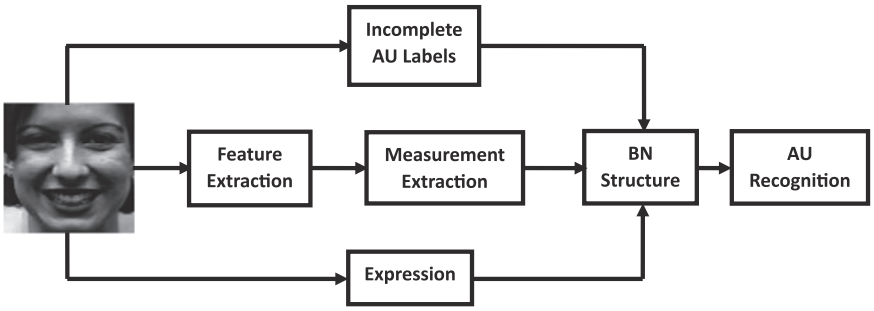


Fig. 1. The framework of our method.

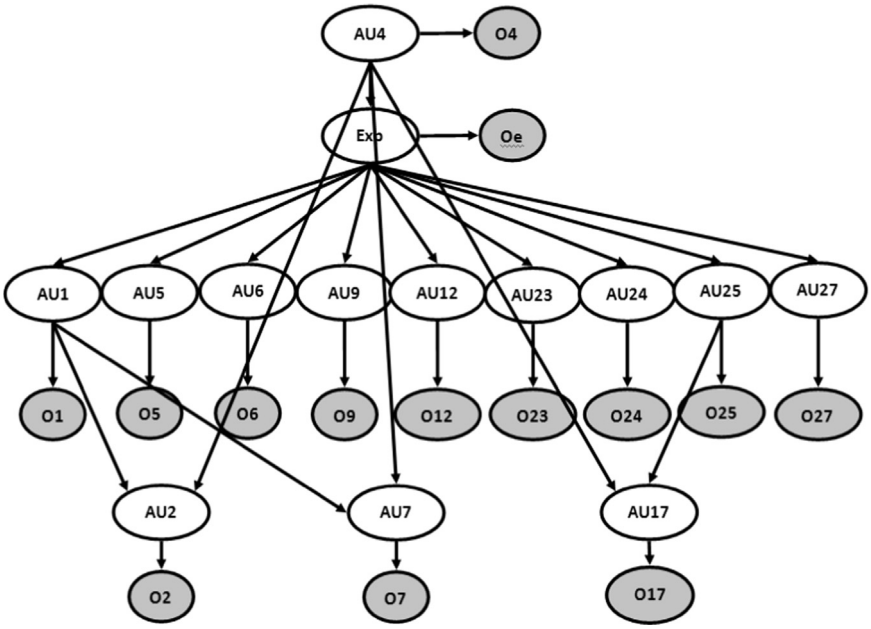


Fig. 2. The BN model for expression and AUs on the CK+ database. (The unshaded nodes are the hidden nodes we want to infer and the shaded nodes are the corresponding measurement obtained by a traditional image-driven method.)

$$\begin{aligned} &\text{MAX} \quad L_D(\theta) \\ &\text{S. T.} \quad g_{il}(\theta) = \sum_{k=1}^{s_i} \theta_{ilk} - 1 = 0 \end{aligned} \tag{3}$$

where g_{il} imposes the constraint that the parameters of each node sums to 1 over all the states of that node. Solving the above equations, we can get $\theta_{ilk} = \frac{w_{ilk}}{\sum_k w_{ilk}}$.

When AU labels are incomplete, the above structural learning algorithm cannot be used directly, therefore, the SEM algorithm is adopted to address incomplete labels. In SEM, we iterate over a pair of steps until convergence, i.e. model likelihood raises to be 1. In the E-step, we use the current model to generate a completed data set, based on which we compute expected sufficient statistics. In the M-step, we use these expected sufficient statistics to improve our model, including both parameters and structure using the above structural learning algorithm. The detailed learning algorithm is summarized in Algorithm 1.

Algorithm 1. The SEM algorithm for structure and parameter learning of the BN model.

1. Initialize the structure G^0 and the parameter set $\theta^0 = \{P(pa(\lambda_i)|\lambda_i), P(\lambda_i)\}$, ($pa(\lambda_i)$ is the parent of λ_i) for the BN model-based on the incomplete training data, in which the missing AU elements are regarded as absence for the initialization in this step. This structure is initialized with expression label and AU labels from the complete portion of the dataset using an existing BN learning algorithm [29].

repeat

2. E-step Using the new BN structure and parameters to infer the missing labels of the incomplete samples and update the training set D.
3. M-step

$$\{G^t, \theta^t\} = \underset{G,\theta}{\text{argmax}} Q(G, \theta: G^{t-1}, \theta^{t-1}) \tag{4}$$

In which, $Q(G, \theta: G^\star, \theta^\star)$ is the expectation of the Bayesian information criterion (BIC) score of any BN $\{G, \theta\}$ calculated using a distribution of the data $P(D|G^\star, \theta^\star)$, which is defined as following:

$$Q^{BIC}(G, \theta: G^*, \theta^*) = E_{G^*,\theta^*}[\log P(D|G, \theta)] - \frac{Dim(G)}{2} \log N \tag{5}$$

where the first term is the log-likelihood function of structure G with respect to data D, representing how well G fits the data. The second term is a penalty relating to the complexity of the network. N represents the number of samples in the training set. The equation of $Dim(G)$ is given by:

$$Dim(G) = \sum_{i=1}^{n} |Pa_i \parallel X_i - 1| \tag{6}$$

where n is the number of nodes in the graph G, $|X_i|$ is the number of possible values of node i, and $|Pa_i|$ is the number of possible values of parent node of node i.

until converges

After the BN structure and parameter learning among the target labels, each of the node is connected with its measurement node. Using the ground truth of target labels and their measurements obtained by a traditional image-driven method, it is easy to get the conditional probability distribution (CPD) for each node. Let λ_i and $\hat{\lambda}_i$, $i \in \{1, \ldots, n, n+1\}$, respectively denote the label variable and the corresponding measurement, the CPD of node i is represented as $P(\hat{\lambda}_i|\lambda_i)$.

3.2.2. BN inference

During the BN inference, the posterior probability of categories can be estimated by combining the likelihood from measurement with the prior model. Most probable explanation (MPE) [30] inference is used to estimate the joint probability.

$$Y^{\star} = \arg\max_{\lambda_1,\ldots,\lambda_n} P(\lambda_1, \ldots, \lambda_n | \hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\lambda}_{n+1})$$
$$= \arg\max_{\lambda_1,\ldots,\lambda_n} \sum_{\lambda_{n+1}} \left(\prod_{i=1}^{n+1} P(\hat{\lambda}_i|\lambda_i) \prod_{i=1}^{n+1} P(\lambda_i|pa(\lambda_i)) \right) \quad (7)$$

where $pa(\lambda_i)$ is the parent of λ_i. The condition probabilities in the equation are learned from training set. In the work, the inferred labels are assigned to be the values $(\lambda_1, \ldots, \lambda_n)$ with the highest probability given $(\hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\lambda}_{n+1})$.

4. Experiments

4.1. Experimental conditions

Experiments of both AU recognition and AU intensity estimation are conducted to validate our method.

4.1.1. AU recognition

For AU recognition, the extended Cohn–Kanade dataset (CK+) [3] which provides 7 expression categories (i.e. anger, contempt, disgust, fear, happy, sadness and surprise), 30 AU labels and 68 normalized vertex points (as shown in Fig. 3(a)) for its samples is adopted. The displacement of 68 normalized landmark points between the apex frame and the onset frame is used as a feature vector. Finally, 327 posed samples with both expression category and FACS labels are selected, and 13 AUs whose frequencies of all the selected samples are more than 10% are considered, which are: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU17, AU23, AU24, AU25, and AU27.

To validate the supplementary role of the expression in assisting AU recognition, three experiments are conducted: the image-driven AU recognition, the model-based method that models only AU relations and our proposed method. Our proposed method considers both the relations among AUs, and the relations among AUs and expressions. The image-driven AU recognition is the same as the initial AU measurement extraction discussed in Section 3.1.

To illuminate the effectiveness of our constructed AU classifier under incomplete AU annotation, we miss each AU label at random with certain probabilities, i.e. 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50%. We compare our proposed method with image-driven method, and the model-based method that models only AU relations. Both our proposed method and model-based method considering only AU relations handle incomplete labels by SEM, while the image-driven method are trained with only the complete portion of the data. 10-fold cross-validation is adopted.

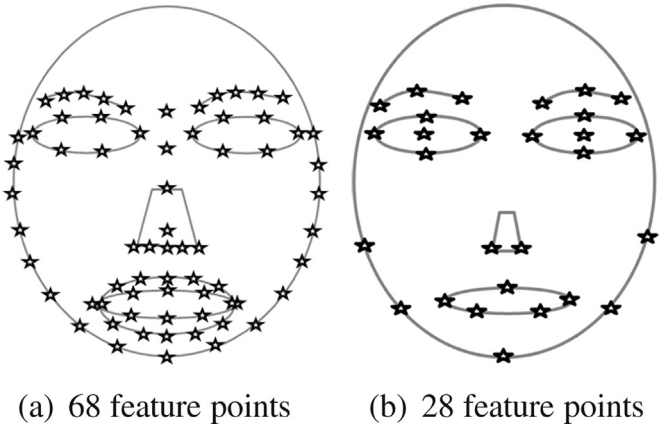


Fig. 3. The schematic diagram of the feature points labeled on the face.

Furthermore, in order to evaluate the generalization ability of the proposed method, the cross-database AU recognition experiments are conducted. We train the BN models using the AU and expression labels from the CK+ database, and then combine the learned BN model and the obtained measurements from the ISL database [31]. The ISL database is collected under real-world conditions with uncontrolled illumination and background, as well as moderate head motion. The 19 frontal view video clips for 7 subjects displaying facial expressions are adopted in this work. For this database, we manually selected onset and apex frame for each clip, and labeled 28 feature points on each selected frames using Tong et al. 's [32] algorithm, as shown in Fig. 3(b). Similarly, the normalized vertex point's movement between the apex frame and onset frame is used as feature vector.

AU recognition is a multi-label classification problem, whose evaluation metric is different from that of single label classification, since for each instance there are multiple labels which may be classified as partly correct or partly incorrect. Thus, there are two kinds of commonly used metrics, example-based and label-based measures [33], evaluating the multi-label classification performance from the view of instances and labels respectively. We adopt both metrics in this work. Let Y_j^i denote the ground truth for the ith label of instance j, which is a binary vector, and Z_j^i is the predicted value for the ith label of instances j, m represents the number of the instances and n is the number of labels.

The example-based measure Hamming Loss, and the label-based measure averaged F_1 score, are defined in Eqs. (8) and (9) respectively. Hamming Loss calculates the fraction of the wrong labels to the total number of labels. It is a loss function and it is upper bounded by the 0–1 loss function. Thus the lower Hamming Loss represents the better performance. F1 score considers both the precision and the recall to evaluate the performance. F1 score reaches its best value at 1 and worst at 0. We expect that higher F1 score yields better performance.

$$HammimgLoss = \frac{\sum_{j=1}^{m} xor(Y_j, Z_j)}{m \times n} \quad (8)$$

$$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \sum_{j=1}^{m} Y_j^i Z_j^i}{\sum_{j=1}^{m} Y_j^i + \sum_{j=1}^{m} Z_j^i} \quad (9)$$

4.1.2. AU intensity estimation

For AU intensity estimation, we conduct experiments on the BP4D-Spontaneous database [34]. The database consists of 328 videos which are captured from 41 subjects. Each subject is asked to attend 8 emotion-elicitation experiments. 5 AUs (AU6, AU10, AU12, AU14 and AU17) are coded with intensities ranged from 0 to 5. FERA 2015 [23] has provided baseline results for the AU intensity estimation task on the BP4D database, and we adopt the 316-dimensional geometric features and experimental settings mentioned in [23] for a fair comparison. The database is divided into 2 partitions. The training partition contains samples from 21 subjects, i.e. F001, F003, F005, F007, F009, F011, F013, F015, F017, F019, F021, F023, M001, M003, M005, M007, M009, M011, M013, M015 and M017. The test partition contains the other 20 subjects, i.e. F002, F004, F006, F008, F010, F012, F014, F016, F018, F020, F022, M002, M004, M006, M008, M010, M012, M014, M016 and M018.

The training samples used at measurement extraction step and BN training step are different in our experiments. We need expression labels at BN training step, so we adopt apex frames as samples. While at the AU measurement extraction step, the expression labels are not required. We define the apex frames of a video sequence as the frames with largest sum of AU intensity

values and then pick out the apex frames from each video sequence as experimental samples, resulting in 2993 samples in the training partition for BN training and 3561 samples in the test partition. As to the samples that are used at the measurement extraction step, we use all the frames in the database instead of adopting the same samples as [23]. Their training samples are subsampled from all the frames to approach a balanced number of positive and negative samples, but they do not give a clear statement about the proportion of sampling and the exact number of the samples of each AU. Furthermore, it is hard to balance the number of samples corresponding to each possible intensity value because the higher intensity values appear much less than the lower ones. If we force balancing the samples, the number of samples will decrease dramatically.

Similar to AU recognition experiments, to validate the supplementary role of expression categories in assisting AU intensity estimation, three experiments are conducted: the image-driven AU intensity estimation, the model-based method that models only AU relations and our proposed method. The image-driven AU intensity estimation is the same as the initial AU measurement estimation discussed in Section 3.1.

To illuminate the effectiveness of our constructed AU intensity estimation under incomplete AU annotation, we leave out each AU intensity at random with 10 varying proportions, i.e. from 5% to 50%. We compare our proposed method with the image-driven method, and the model-based method that models only AU relations. Both our proposed method and the model-based method handle incomplete labels with the SEM, while the image-driven method is trained with the incomplete data. We adopt Pearson correlation coefficient (PCC) and intraclass correlation coefficient (ICC) as measures in order to compare our results with the baseline ones in [23]. The mean square error (MSE) is not adopted because [23] views the AU intensity estimation task as a regression problem while we see it as a classification problem. The equations of adopted measures are shown as follows:

$$PCC = \frac{\sum_{j=1}^{m}(Z_j - \overline{Z})(Y_j - \overline{Y})}{(n-1)\sigma_Z \sigma_Y} \tag{10}$$

$$ICC = \frac{\sum_{j=1}^{m}(Z_j - \overline{(Z,Y)})(Y_j - \overline{(Z,Y)})}{(n-1)\sigma^2_{(Z,Y)}} \tag{11}$$

PCC measures the linear correlation between two vectors $\vec{x}$ and $\vec{y}$ with a value between 0 and 1. The value 1 represents total positive correlation and -1 represents total negative correlation. ICC describes how strongly variables in the same group resemble each other. The value of ICC ranges from 0 to 1. We expect an ICC value that is greater than 0.6 for a reliable result. Since PCC only measures the linear correlation, ICC can be a good supplement for it.

4.2. Experimental results of AU recognition

Section 4.2 is organized as follows: Section 4.2.1 focuses on analyzing the dependencies modeling by the BN. The experimental results are shown in Sections 4.2.2 and 4.2.3. Section 4.2.2 analyzes the results under complete data, and Section 4.2.3 analyzes the ones under incomplete data. The comparison with related work is discussed in Section 4.2.4. In Section 4.2.5, we show the results of cross-database experiments.

4.2.1. Results and analysis of AUs and expressions dependencies modeling by the BN on the CK+ database

We quantify the dependencies among different AUs and the dependencies among AUs and expressions using a conditional probability of $P(\lambda_j|\lambda_i)$, as shown in Table 1, which measures the probability of label λ_j happens, given label λ_i happens. From Table 1, we can find that there exist two kinds of relationships among AUs, and among AUs and expressions: co-occurrence and mutual exclusion. For example, $P(AU1|AU2)$ is 1.00, which shows AU1 is always coexistent with AU2. $P(AU25|surprise)$ is 0.988, indicating that AU25 is an important AU to express surprise. $P(AU1|AU9)$ and $P(AU1|happy)$ are 0.00, which means AU1 never coexists with AU9, and AU1 is rarely active when users express happiness. The proportions of the co-occurrence and mutual exclusion relations among AUs and expressions are larger than those among AUs. Specifically, 39.6% and 22.0% relations among AUs and expressions are extremely mutually exclusive (i.e. $P(\lambda_j|\lambda_i) = 0.00$) and co-occurrence (i.e. $P(\lambda_j|\lambda_i) > 0.70$) respectively while there is only 14.7% relations among AUs are extremely mutual exclusion

Table 1
Dependencies among labels on the CK+ database (each entry a_{ij} represents $P(\lambda_j = 1|\lambda_i = 1)$).

λ_i	λ_j												
	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU12	AU17	AU23	AU24	AU25	AU27
AU1	Null	**0.762**	0.315	0.646	0.023	0.038	**0.000**	0.038	0.223	0.038	0.008	**0.777**	0.546
AU2	**1.000**	Null	0.121	**0.788**	**0.000**	**0.000**	**0.000**	0.040	0.101	0.030	0.010	**0.909**	**0.717**
AU4	0.336	0.098	Null	0.156	0.180	0.508	0.311	0.025	**0.754**	0.287	0.254	0.213	0.008
AU5	**0.913**	**0.848**	0.207	Null	0.033	0.043	0.011	0.043	0.098	0.076	**0.000**	**0.924**	0.652
AU6	0.032	**0.000**	0.232	0.032	Null	0.347	0.189	**0.705**	0.211	0.074	0.032	**0.737**	**0.000**
AU7	0.063	**0.000**	**0.785**	0.051	0.418	Null	0.443	0.076	0.696	0.329	0.329	0.228	**0.000**
AU9	**0.000**	**0.000**	0.623	0.016	0.295	0.574	Null	0.033	**0.705**	0.082	0.148	0.148	**0.000**
AU12	0.063	0.050	0.038	0.050	**0.838**	0.075	0.025	Null	0.025	**0.000**	0.025	**0.900**	0.038
AU17	0.252	0.087	**0.800**	0.078	0.174	0.478	0.374	0.017	Null	0.304	0.278	0.052	**0.000**
AU23	0.116	0.070	**0.814**	0.163	0.163	0.605	0.116	**0.000**	**0.814**	Null	0.581	0.023	0.023
AU24	0.023	0.023	**0.721**	**0.000**	0.070	0.605	0.209	0.047	**0.744**	0.581	Null	**0.000**	**0.000**
AU25	0.558	0.497	0.144	0.470	0.387	0.099	0.050	0.398	0.033	0.006	**0.000**	Null	0.398
AU27	**0.986**	**0.986**	0.014	**0.833**	**0.000**	**0.000**	**0.000**	0.042	**0.000**	0.014	**0.000**	**1.000**	Null
Anger	**0.000**	**0.000**	**0.889**	0.133	0.178	**0.711**	0.067	0.022	**0.867**	**0.800**	**0.733**	**0.000**	**0.000**
Contempt	0.056	0.056	0.056	**0.000**	**0.000**	**0.000**	**0.000**	0.278	0.278	0.056	0.111	**0.000**	**0.000**
Disgust	**0.000**	**0.000**	0.610	**0.000**	0.305	0.559	**0.983**	0.034	0.695	0.034	0.119	0.153	**0.000**
Fear	**0.880**	0.400	**0.840**	0.640	0.120	0.240	**0.000**	0.080	0.120	**0.000**	**0.000**	**0.920**	**0.000**
Happy	**0.000**	**0.000**	**0.000**	**0.000**	**0.957**	0.101	**0.000**	**0.971**	**0.000**	**0.000**	**0.000**	**0.971**	**0.000**
Sad	**0.929**	0.250	**0.821**	**0.000**	**0.000**	0.036	**0.000**	**0.000**	**0.964**	0.107	0.036	**0.000**	**0.000**
Surprise	**0.976**	**0.976**	0.012	**0.843**	**0.000**	**0.000**	**0.000**	0.036	**0.000**	0.012	**0.000**	**0.988**	**0.867**

The bold values indicate the co-current ($P(\lambda_j|\lambda_i) > 0.70$) and extremely mutually exclusive ($P(\lambda_j|\lambda_i) = 0.00$) relation.

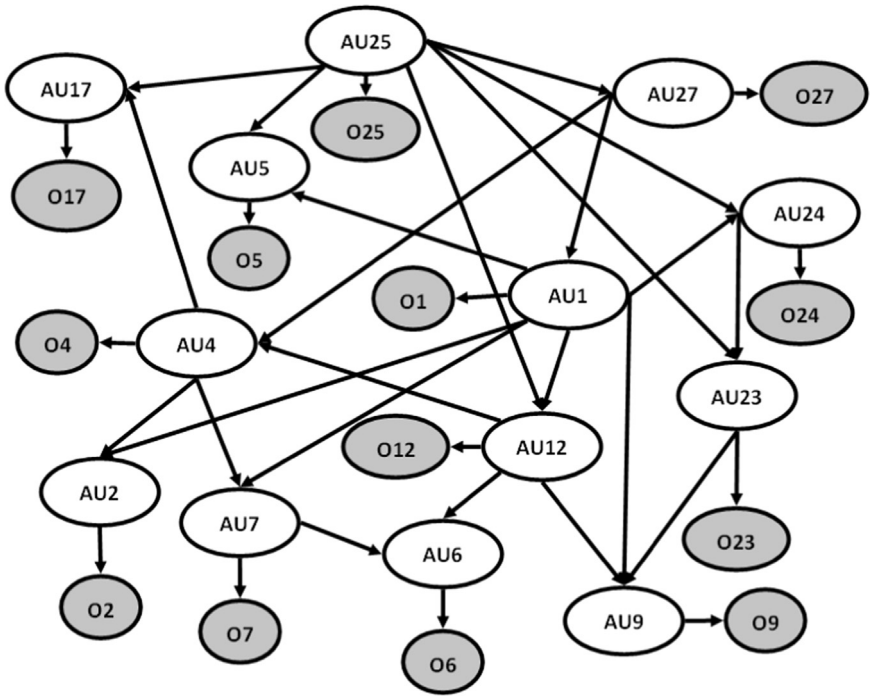


Fig. 4. The BN model for AUs on the CK+ database. (The unshaded nodes are the hidden nodes we want to infer and the shaded nodes are the corresponding measurement obtained by a traditional image-driven method.)

and co-occurrence. It indicates that the relations among AUs and expressions may be stronger that those among AUs.

To systematically capture dependencies among AUs, and dependencies among AUs and expressions, we learned two BNs, shown in Fig. 2 and Fig. 4 respectively. In Fig. 2, expression variable takes 7 values representing 7 types of expression. The links in the structure represent the dependencies among labels. Comparing two learned BNs with the dependency tables, we find that the label pairs whose conditional probabilities are top ranked or bottom ranked are linked in the BNs in most cases. It demonstrates the effectiveness of the BN structure learning method which can effectively capture the mutually exclusive and co-existent relationships among multiple labels. Comparing Fig. 2 with Fig. 4, we find that by adding the expression node, many links among AUs are removed. Specifically, in Fig. 4, each AU node is connected with at least two other AU nodes, while in Fig. 2, all of the 13 AUs nodes are connected with the expression node directly or indirectly and most AUs are conditionally independent given the expression. It further confirms that the relations among AUs and expressions are stronger than those among AUs. It also says that AU relations are mostly expression-dependent or AUs are more likely to relate to each other via the expression. Specifically, AU dependencies can be classified into two types: expression-dependent and expression-independent. While links among AUs capture the direct relationships among AUs, AUs linked through the expression node capture the expression-dependent AU relationships. For Fig. 4, the two types of AU relationships are all captured by the direct links among AUs. As a result, the model is denser. While in Fig. 2, by introducing the expression node, many inherent links among AUs disappear. This means AUs are mostly dependent on each other through the facial expression, i.e. most of the AU relationships are expression-dependent. The remaining direct links among AUs in Fig. 2 capture the expression-independent AU dependencies, i.e. AU dependencies as a result of the underlying facial anatomy, taking the link between AU1 and AU2 for example, both AU1 and AU2 are related to frontalis facial muscle [35]. Therefore, the AU recognition performance may be better improved by considering both the relations among AUs, and the relations among AUs and expressions.

Table 2
Experimental results with complete AU annotations.

Method	CK+		Cross-database		BP4D	
	Hamming	F_1 score	Hamming	F_1 score	PCC	ICC
Image-driven	0.1006	0.738	0.105	0.834	0.552	0.603
Model-based	0.100	0.793	0.101	0.844	0.576	0.617
Proposed	**0.096**	**0.800**	**0.097**	**0.846**	**0.600**	**0.632**

4.2.2. Results and analysis of AU recognition with complete AU annotations on the CK+ database

The experimental results of AU recognition are shown in Table 2. From the table, we can obtain the following observations:

1. The image-driven method performs worst among the three methods, since its Hamming Loss is the highest, and the F_1 score is the lowest. The image-driven method predicts each AU independently from the image features only. However, the AUs are not totally independent. There exist co-occurrent and mutually exclusive relations among AUs, and the relations among AUs and expressions are even much stronger as discussed above. The two learned BNs systematically capture these relations. Our AU recognition methods capturing expression-dependent AU relations and combining it with AU measurements improve the performance.
2. The AU recognition by considering both AUs' relations and AU-expression relations performs better than the model-based method only considering AU relations, with lower Hamming Loss and higher F_1 score. It suggests that the stronger relations among AUs and expressions can facilitate AU recognition better than using the AU relations only, even when the expression labels are only used during training.

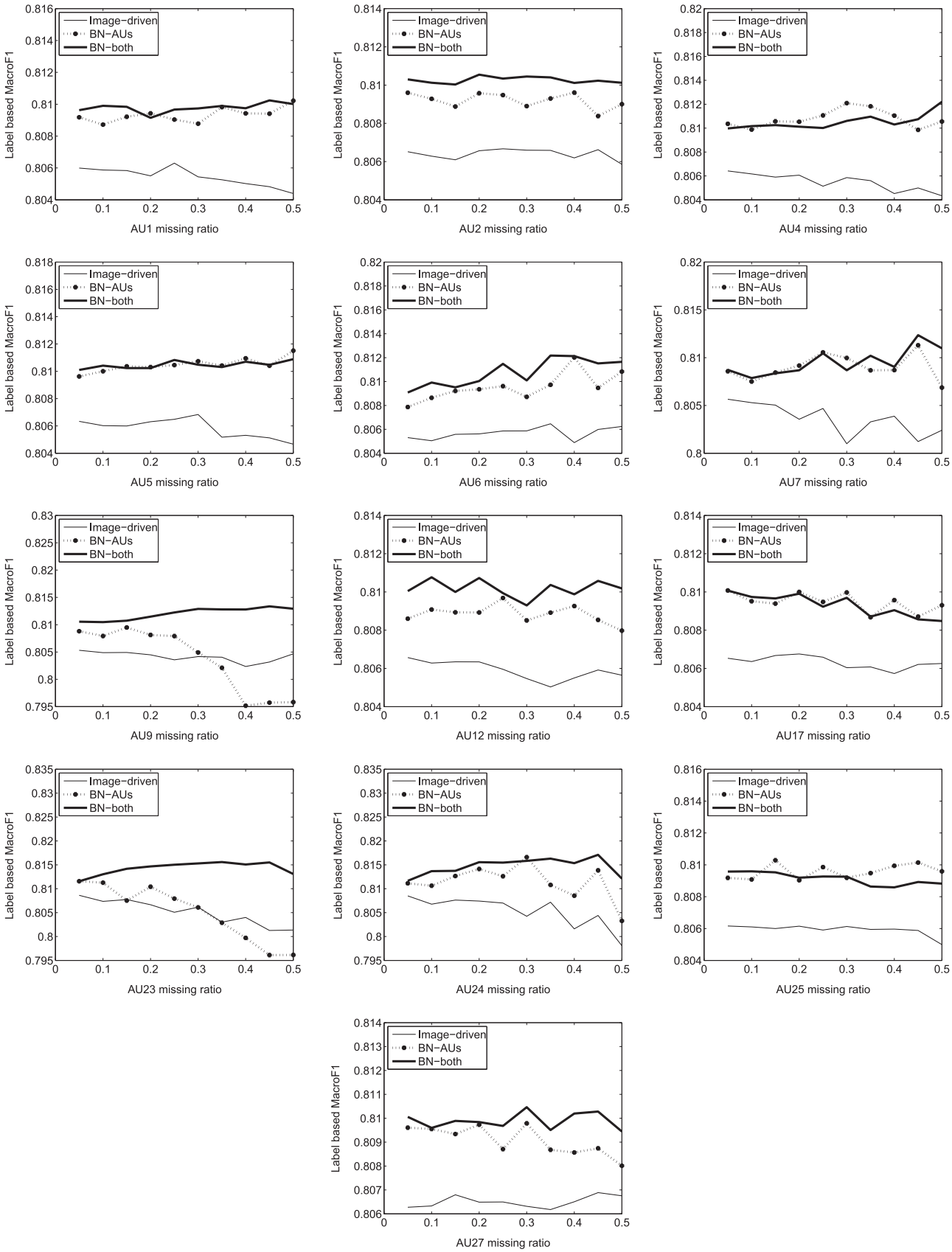


Fig. 5. AU recognition results on CK+ database under 10 different missing proportions.

To demonstrate that the performance improvement is statistically significant, we conducted 10-fold cross-validation 10 times and a Wilcoxon signed-rank test. The p-value of model-based vs. image-driven, ours vs. image-driven, ours vs. model-based in terms of Hamming Loss and F_1 with complete AU labels are all less than 0.05, showing that the performance improvement is statistically significant.

4.2.3. Experimental results with incomplete AU annotations on the CK+ database

To investigate robustness of the proposed method with respect to missing AUs, we conducted the 10-fold cross-validation seven times, and the average results are summarized in Fig. 5. From Fig. 5, we can conclude the following:

1. The image-driven method performs worst among the three methods in most cases, since it is trained based complete AU data and ignores incomplete AU data. It indicates that the learned BN structure and proposed EM algorithm can effectively handle missing labels.
2. Our method considering both AUs' relations and AU-expression relations outperforms the model-based AU recognition considering only AUs relations in eight cases, i.e. AU1, AU2, AU6, AU9, AU12, AU23, AU24 and AU27. The performance of the former is almost similar as that of the latter on three AUs, including AU5, AU7 and AU17. For AU4 and AU25, the performance of the former is slightly weaker than that of the latter. Therefore, in general, expressions can better help AU recognition with incomplete AU annotation, due to the stronger relations among AUs and expressions.
3. The performances of AU9 and AU23 degrade severely with the model-based method when the missing ratio increases. We analyze the experimental data and find that the AU9 and AU23 are the least present two AUs. Among all the 327 samples of the CK+ database, AU9 is present in only 61 samples and AU23 is present in only 43 samples. We consider that this imbalance leads to the biased dependency among AUs modeled by the model-based method and hence unsatisfactory performance for AU9 and AU23. When our method is performed, the expression information can compensate for this imbalance and lead to the much better results than both of the model-based method and the image-driven method.

4.2.4. Comparison with related AU recognition work

Although many studies have been done on AU recognition, and achieved good performance, only a small part of the studies exploit the dependencies among AUs as described in Section 2. Among the three model-based works, Tong et al. [36] conducted the experiments on the CK database. Since they do not report the experimental results on the CK+ database, we cannot compare with their work directly. However, the AU recognition considering AU relations only is similar to Tong et al.'s [36] work. From Section 4.2.2, we can find that our proposed method using expression to assist AU recognition performs better than Tong et al.'s [36] work, when the AU annotation is complete. The other two model-based works, Wang et al. [16] and Eleftheriadis et al. [15] validated their proposed methods on the CK+ database, therefore, we list their results in Table 3. In Wang et al.'s [16] work, the results corresponding to each AU are plotted onto the figures and not listed numerically. Therefore, we just compare the average result. From Table 3 we can find that our average result outperforms the two reference works.

There exists lots of image-driven AU recognition works that evaluate on the CK+ databases. Since the used samples (all samples [37] or part of samples [27,38]), the recognized AUs (8 AUs [37], 9 AUs [38] and 13 AUs [27]), and the validation strategy (such as n-fold cross-validation [37,38] or leave-one-subject-out [27]) vary with methods, it is hard to compare our results with theirs fairly. We only compare the experimental results as a reference. Furthermore, we list current fixed AU combination works [11,12] in Table 4. Table 4 reports the comparison of the proposed method with the reference works, which also used F_1 score as the evaluation metric. From Table 4 we can find that among 13 AUs, the results of 5 AUs are the best among the reference works. Although the average F_1 scores of [38,11] are higher than ours, [38] only consider 7 AUs and [11] considers 9 AUs, while our work considers 13 AUs. The average F_1 score of such 7 AUs using our method is 0.868, which is higher than [38]. If we calculate the average F_1 score under the same condition with [11], our result is 2.4% higher than theirs. It shows that for complete AU annotation, our method outperforms the current image-driven method.

For comparison under incomplete AU annotation, we performed two comparisons. First, we compare the model-based method considering only AU relations with our method modeling both AU relations as well as AU and expression relations trained using only the complete portions of the training data. Second, we compare the two methods using all training data. The average results of the first study are summarized in Table 5, which again shows that our model outperforms Tong's model. It demonstrates once again that adding the expression label can improve AU

Table 3
Comparison with current model-based method on the CK+ database (in F_1 score).

AUs	Proposed	[16]	[15]
1	**0.945**	–	0.825
2	**0.935**	–	0.870
4	0.785	–	**0.792**
5	**0.798**	–	0.735
6	**0.776**	–	0.728
7	**0.601**	–	0.575
9	**0.911**	–	0.879
12	0.847	–	**0.875**
17	0.837	–	**0.868**
23	**0.689**	–	0.673
24	0.447	–	**0.510**
25	**0.952**	–	0.918
27	**0.977**	–	0.911
Average	**0.800**	0.792	0.781

The bold values denote the best experimental results among methods listed in a Table.

Table 4
Comparison with current image-driven AU recognition works on the CK+ database (in F_1 score).

AUs	Proposed	[27]	[37]	[38]	[11]	[12]
1	**0.945**	0.779	0.622	0.88	0.91	0.900
2	**0.935**	0.801	0.762	0.92	0.88	0.930
4	0.785	0.775	0.691	0.89	**0.90**	–
5	**0.798**	0.636	–	–	0.74	–
6	0.776	0.771	0.796	**0.93**	0.91	0.742
7	0.601	0.624	**0.791**	–	–	0.667
9	**0.911**	0.788	–	–	0.82	–
12	0.847	**0.900**	0.772	0.90	0.90	0.807
17	0.837	0.811	**0.843**	0.76	0.75	0.835
23	0.689	–	–	–	–	**0.743**
24	0.447	0.601	–	–	–	**0.659**
25	**0.952**	0.889	–	0.73	0.76	–
27	0.977	**0.995**	–	–	–	–
Average	0.800	0.780	0.754	**0.859**	0.841	0.785

The bold values denote the best experimental results among methods listed in a Table.

Table 5
Comparison experimental results on the CK+ database under 10 different missing proportions ($_{*}10^{-1}$).

Mis			5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
H	*Im*	*Ave*	0.875	0.877	0.877	0.877	0.877	0.879	0.880	0.881	0.881	0.882
		Std	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
	B_1	*Ave*	0.860	0.860	0.861	0.860	0.858	0.862	0.860	0.865	0.862	0.865
		Std	0.00	0.01	0.01	0.01	**0.00**	0.01	0.00	0.01	0.01	0.01
	B_2	*Ave*	**0.855**	**0.854**	**0.856**	**0.854**	**0.855**	**0.857**	**0.856**	**0.860**	**0.857**	**0.859**
		Std	0.00	**0.00**	**0.00**	0.01	0.01	0.01	0.00	0.01	0.01	0.01
F	*Im*	*Ave*	8.065	8.061	8.062	8.060	8.058	8.054	8.054	8.047	8.048	8.043
		Std	0.04	0.02	0.03	0.04	0.03	0.05	0.04	0.05	0.07	0.08
	B_1	*Ave*	8.097	8.096	8.095	8.097	8.097	8.090	8.093	8.078	8.084	8.080
		Std	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	**0.02**	0.03
	B_2	*Ave*	**8.103**	**8.104**	**8.100**	**8.106**	**8.103**	**8.098**	**8.102**	**8.091**	**8.098**	**8.093**
		Std	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.03	0.03

"Mis" represents "missing proportions", "H" represents "Hamming Loss", " *F*" represents "F_1 score", "Im" represents "Image-driven", "B_1" represents "BN structure considering only AUs without EM algorithm", "B_2" represents "BN structure considering both AUs and expression without EM algorithm", "Ave" represents "average", and "Std" represents "standard error".
The bold values denote the best experimental results among methods listed in a Table.

recognition even under insufficient fully labeled AU data. For the second study, since Tong's model cannot handle incomplete data, we extend Tong's model to handle incomplete data using EM algorithm. From Section 4.2.3, we can find that our method considering both AUs' relations and AU-expression relations outperform the model-based AU recognition considering only AUs relations when the AU annotation is incomplete. Thus, our method outperforms the current model-based AU recognition for both complete and incomplete AU annotations.

4.2.5. Cross-database AU recognition experiments

Tables 2 and 6 show the cross-database AU recognition results under complete and incomplete AU annotations respectively. From the two tables, we can find that the performance of the proposed method is better than that of the model-based method considering AU relations only for both complete and incomplete AU annotations, with lower hamming distance and higher F_1 score. Our method performs more stably than the model-based method considering AU relations only, since our method has the lower standard deviation of the two parameters. Therefore, the cross-database experiments demonstrate the generalization ability and the robustness of our method.

4.3. Experimental results of AU intensity estimation

Section 4.3 is organized as follows: Section 4.3.1 focuses on analyzing the dependencies modeling by the BN. The experimental results are shown in Sections 4.3.2 and 4.3.3. Section 4.3.2 analyzes the results under complete data and Section 4.3.3 analyzes the ones under incomplete data. The comparison with related work is discussed in Section 4.3.4

4.3.1. Results and analysis of AU intensity and expression dependencies modeling by a BN

After validating the effectiveness of our method on AU occurrence data, we extend the method to the AU intensity estimation task. To systematically capture relationships among AU intensities, and relations among AU intensities and the expression labels, i.e. the expressions on the BP4D database, we have learned 2 BNs that are shown in Fig. 6.

From Fig. 6 we can obtain the similar observations to those on the CK+ database. There are totally six directed edges among AU nodes in the AUs' model, while after adding expression node to the model, about half of the directed edges are replaced by those of the expression-dependent relations. It further confirms that the relations between expressions and AUs are stronger than those among AUs. From Fig. 6(b), we can find that the expression node is connected to 4 out of 5 AUs, which shows strong relationship among expressions and AUs.

4.3.2. Results and analysis of AU intensity estimation with complete annotations

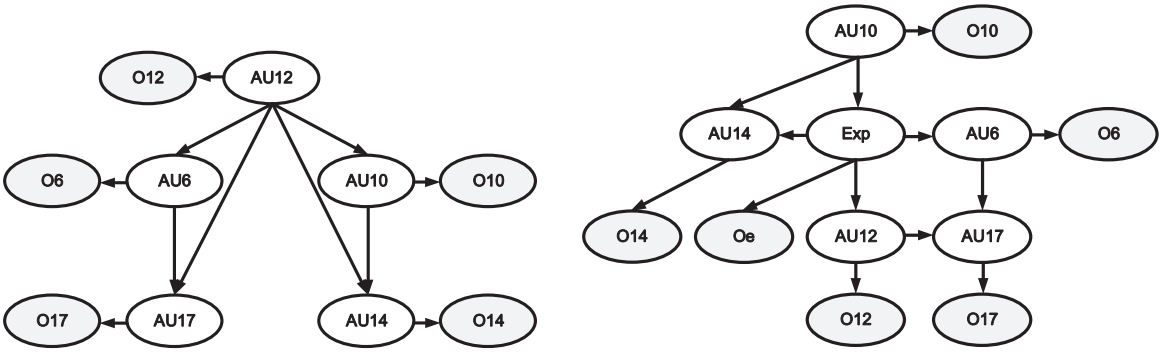
We perform experiments with complete AU labels on the BP4D database 10 times, then we take the average results in order to avoid the errors caused by randomly initial parameters. We calculate the variance of the 10 results and it is of the order of $1e-2$, which demonstrates the stability of our method. The experimental results are shown in Table 2.

From the BP4D part of Table 2, we can conclude that for AU intensity estimation, our method outperforms both the image-

Table 6
Average of 13 AUs recognition results on the ISL database under different missing proportions ($_{*}10^{-1}$).

Mis			5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
H	*M*	*Ave*	1.004	0.995	0.997	1.000	1.022	1.011	1.019	0.997	1.032	1.022
		Std	0.03	0.06	0.05	0.06	0.04	0.04	0.06	0.08	0.05	0.05
	P	*Ave*	**0.975**	**0.984**	**0.978**	**0.985**	**0.988**	**0.976**	**0.995**	**0.980**	**0.975**	**0.979**
		Std	**0.01**	**0.02**	**0.02**	**0.04**	0.04	**0.02**	**0.05**	**0.03**	**0.04**	**0.03**
F	*M*	*Ave*	8.434	8.427	8.404	8.404	8.360	8.375	8.342	8.386	8.330	8.340
		Std	0.04	0.11	0.10	0.07	0.09	0.08	0.13	0.09	0.09	0.10
	P	*Ave*	**8.453**	**8.437**	**8.448**	**8.439**	**8.416**	**8.439**	**8.410**	**8.444**	**8.431**	**8.434**
		Std	**0.02**	**0.04**	**0.04**	**0.06**	**0.08**	**0.04**	**0.08**	**0.06**	**0.08**	**0.05**

"Mis" represents "missing proportions", "H" represents "Hamming Loss", " *F*" represents "F_1 score", " *M*" represents "Model_based method", " *P*" represents "Proposed method", "Ave" represents "average", and "Std" represents "standard error".
The bold values denote the best experimental results among methods listed in a Table.



(a) The BN model for AUs. (b) The BN model for AUs and expressions.

Fig. 6. The BN models on the BP4D database.

Table 7
Comparison with [23] on the BP4D database.

Measure	Method	AU6	AU10	AU12	AU14	AU17	Average
ICC	[23]	0.690	**0.696**	0.653	**0.453**	0.278	0.554
	Proposed	**0.768**	0.662	**0.839**	0.198	**0.530**	**0.600**
PCC	[23]	0.699	**0.715**	0.706	**0.472**	0.365	0.592
	Proposed	**0.781**	0.671	**0.843**	0.289	**0.577**	**0.632**

The bold values denote the best experimental results among methods listed in a Table.

driven and the model-based methods on the BP4D database. It further demonstrates that capturing the dependencies among AUs and expressions instead of capturing the ones among AUs can improve the performance of the model. The dependencies highly supplement AU intensity estimation.

To further validate the improvement of our method to image-driven and model-based methods, we perform Wilcoxon signed-rank test on the 10 group results, and the p-value of ours vs. image-driven and ours vs. model-based in terms of PCC and ICC with complete AU labels are all less than 0.05, showing that the performance improvement is statistically significant.

4.3.3. Comparison with the baseline work on the BP4D database

There exist several works that estimate AU intensity on the BP4D database [23,39–42]. In this part, we focus on comparing our results on the database with [23] because our experimental conditions are much the same. The adopted features in our experiments are provided by Valstar [23]. We also use the same metrics and cross validation partitions to evaluate the performance as [23], which makes the comparison relatively fair. The detailed comparison is shown in Table 7. Combining Tables 7 and 2 we are able to not only compare the experimental results but also make it clear that how our method improves at the basis of the SVR results.

From Tables 7 and 2, we can find that the results corresponding to the image-driven method, the model-based method and our method are gradually increased. The model-based method captures the dependencies among AUs at the basis of image-driven method, which improves the performance. Our method not only captures dependencies among AUs but also captures the ones among AUs and expression, which introduces more useful information to the model. Therefore, our method further improves the performance.

From Table 7 we can conclude that our approach generally outperforms the work [23], since the performance of most AUs and the average results are much better than those of [23]. When comparing to the SVR results referred in [23], we can find that our results of AU6, AU12 and AU17 are better than theirs. Since the experimental conditions are similar, the comparison results are convictive. In other words, our method can outperform [23] under the same experimental conditions.

Since the measurements are important components of our method, the final results still have improvement space if we get better measurement results, which has been confirmed during the parameter adjustment process.

4.3.4. Experimental results with incomplete AU intensity annotations

The results of AU intensity estimation with incomplete AU labels are shown in Fig. 7. From Fig. 7, we can find that our proposed method outperforms both the image-driven method and the model-based method under most of the missing ratios, since the ICC of our method is the highest in most cases, verifying the effectiveness of our proposed method. Compared with the proposed method and model-base method, the performance of image-driven method decreases much faster along with the increase of the missing ratio. It demonstrates that the relations among AUs, as well as the relations among AUs and expressions, are benefit for the robustness of AU intensity estimation.

5. Conclusion

Current AU recognition requires complete AU annotation for training. However, the efforts for training human experts and manually annotating the AUs are expensive and time consuming. Furthermore, the reliability of manual AU annotation is inherently attenuated by the subjectivity of human coder, especially for the AUs that are difficult to label. In contrast, expressions are much easier to annotate with great accuracy. Therefore, we design an AU recognition method with the assistance of expression labels under incomplete AU labeling. We propose to construct a BN model to capture the dependencies not only among AUs but also among AUs and expressions. During training, the image features and expression labels are complete, while the AU labels may be missing. SEM is adopted to learn the structure and parameters of the BN. The traditional image-driven method is adopted to obtain the expression and AU measurements. During testing, the AUs are inferred by combining the measurements and the AU relations in the BN model. Both within database and cross-database experiments, as well as both AU occurrence prediction and AU intensity estimation are conducted to compare our method with image-driven method and model-based method. The experimental results of within database show that our method outperforms the state of art model-based AU recognition methods for both complete and incomplete AU annotations, as well

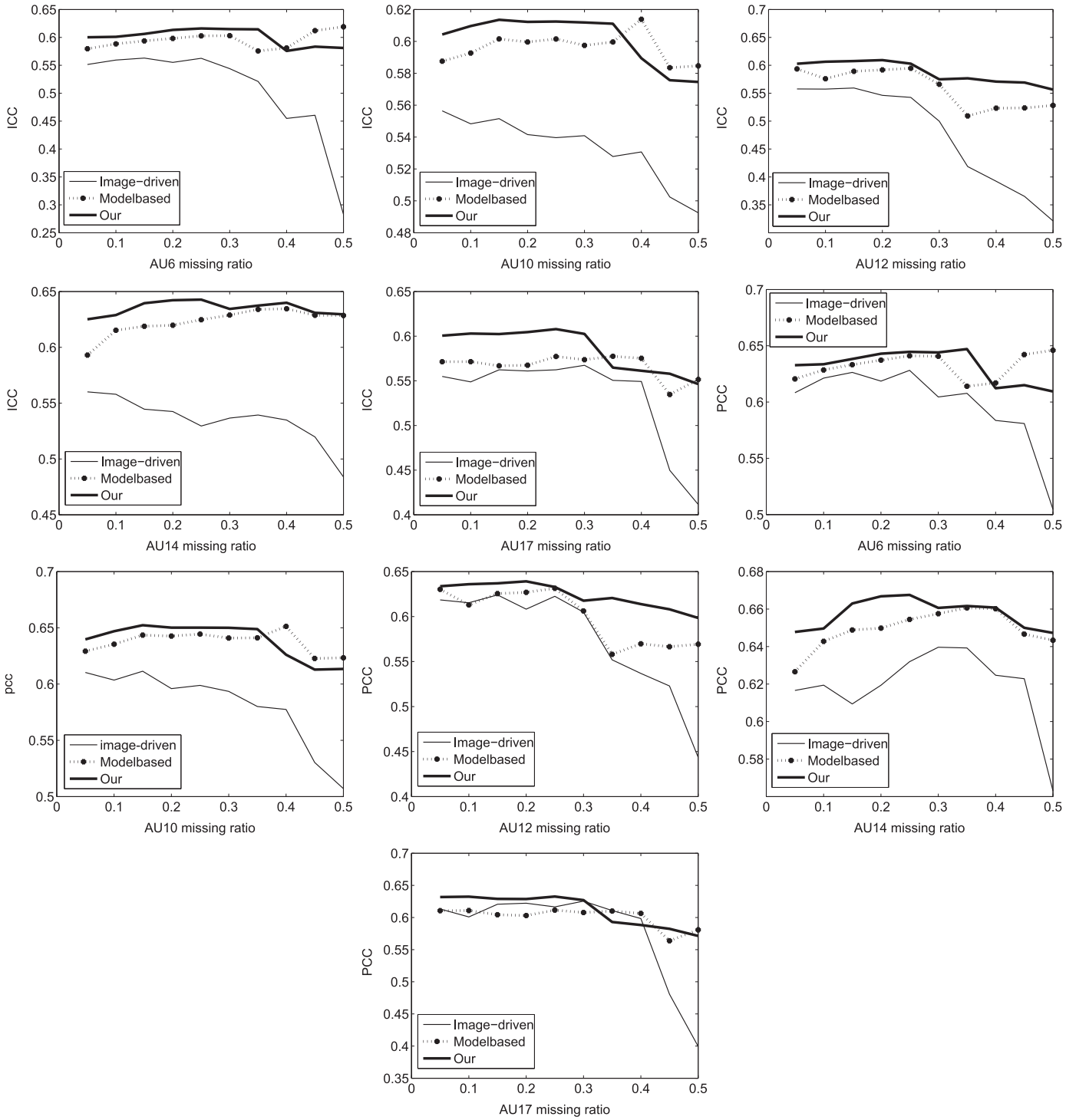


Fig. 7. AU recognition results on the BP4D database under 10 different missing proportions.

as both binary AU occurrences and continuous AU intensities. The AU recognition assisted by expression performs better than the current image-driven method for complete AU annotations. The cross-database experimental results demonstrate that our method has better generalization ability and is more robust compared with current model-based and image-driven methods.

Since the proposed BN is learned from ground truth of AU and expression labels, it can handle not only posed but also spontaneous expressions, assuming that the ground truth of expression labels are representative of the ground truth of AU labels.

The proposed approach is expression-dependent AU recognition. The expressions we study in this work are a small yet the most generic set of prototype expressions. They are universally true in most databases. Our method may not work well for other infrequent expressions. Even AUs can be formed independent of expressions, most AUs are formed in order to perform certain kind of expressions. Moreover, by using the readily available expression during training, we implicitly impose a prior on AUs and hence help improve AU recognition. The prior, of course, can be wrong if the expressions are different from those in the database.

Besides the spatial dependency among AU and expression, the temporary dependency is also crucial for AU analyses. Because of

the need of significance theoretical development, we do not consider the dynamic relationships modeling in this work. We will investigate this issue in the future.

Acknowledgments

This work has been supported by the National Science Foundation of China (grant nos. 61175037, 61228304, and 61473270) and the project from Anhui Science and Technology Agency (1508085SMF223).

References

[1] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[2] P. Ekman, W.V. Friesen, Facial Action Coding system: A Technique for the Measurement of Facial Movement, vol. 12, Consulting Psychologists Press, Palo Alto, CA 1978, pp. 271–302.
[3] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Piscataway, NJ, 2010, pp. 94–101.
[4] A.M. Rahman, M.I. Tanveer, M. Yeasin, A spatio-temporal probabilistic framework for dividing and predicting facial action units, in: Affective Computing and Intelligent Interaction, Springer-Verlag Berlin Heidelberg, 2011, pp. 598–607.
[5] S. Velusamy, H. Kannan, B. Anand, A. Sharma, B. Navathe, A method to infer emotions from facial action units, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Piscataway, NJ, 2011, pp. 2028–2031.
[6] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 699–714.
[7] M.F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (1) (2012) 28–43.
[8] L. van der Maaten, E. Hendriks, Action unit classification using active appearance models and conditional random fields, Cogn. Process. 13 (2) (2012) 507–518.
[9] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, The computer expression recognition toolbox (CERT), in: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE Piscataway, NJ, 2011, pp. 298–305.
[10] S. Lucey, A.B. Ashraf, J. Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, Face Recogn. (2007) 275–286.
[11] X. Zhang, M.H. Mahoor, Simultaneous detection of multiple facial action units via hierarchical task structure learning, in: 2014 22nd International Conference on Pattern Recognition (ICPR), IEEE Piscataway, NJ, 2014, pp. 1863–1868.
[12] K. Zhao, W.-S. Chu, F. De la Torre, J.F. Cohn, H. Zhang, Joint patch and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2207–2216.
[13] Y. Tong, J. Chen, Q. Ji, A unified probabilistic framework for spontaneous facial action modeling and understanding, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 258–273.
[14] Y. Tong, W. Liao, Q. Ji, Inferring facial action units with causal relations, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE Piscataway, NJ, 2006, pp. 1623–1630.
[15] S. Eleftheriadis, O. Rudovic, M. Pantic, Multi-conditional latent variable model for joint facial action unit detection, in: International Conference on Computer Vision (ICCV), Santiago, Chile, 2015.
[16] Z. Wang, Y. Li, S. Wang, Q. Ji, Capturing global semantic relationships for facial action unit recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3304–3311.
[17] M.H. Mahoor, S. Cadavid, D.S. Messinger, J.F. Cohn, A framework for automated measurement of the intensity of non-posed facial action units, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009 (CVPR Workshops 2009) IEEE Piscataway, NJ, 2009, pp. 74–80.
[18] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: 7th International Conference on Automatic Face and Gesture Recognition, 2006 (FGR 2006), 2006, pp. 223–230.
[19] L. Jeni, J. Girard, J. Cohn, F. de la Torre, Continuous Au intensity estimation using localized, sparse facial feature space, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–7. http://dx.doi.org/10.1109/FG.2013.6553808.
[20] A. Savran, B. Sankur, M.T. Bilge, Regression-based intensity estimation of facial action units, Image Vis. Comput. 30 (10) (2012) 774–784, http://dx.doi.org/10.1016/j.imavis.2011.11.008 (3D Facial Behaviour Analysis and Understanding. URL ⟨http://www.sciencedirect.com/science/article/pii/S0262885611001326⟩).
[21] S. Kaltwang, O. Rudovic, M. Pantic, Continuous pain intensity estimation from facial expressions, in: Advances in Visual Computing, Springer, IEEE Piscataway, NJ, 2012, pp. 368–377.
[22] S.W. Chew, Recognising Facial Expressions with Noisy Data, 2013.
[23] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, J. Cohn, in: FERA 2015 – second facial expression recognition and analysis challenge, in: Proceedings of the IEEE ICFG, 2015.
[24] Y. Li, S. Mavadati, M. Mahoor, Q. Ji, A unified probabilistic framework for measuring the intensity of spontaneous facial action units, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–7.
[25] G. Sandbach, S. Zafeiriou, M. Pantic, Markov random field structures for facial action unit intensity estimation, in: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), 2013, pp. 738–745.
[26] A. Ruiz, J. van de Weijer, X. Binefa, From emotions to action units with hidden and semi-hidden-task learning, in: International Conference on Computer Vision (ICCV), 2015 ⟨http://cmtech.upf.edu/research/projects/shtl⟩.
[27] Y. Li, S. Wang, Y. Zhao, Q. Ji, Simultaneous facial feature tracking and facial expression recognition, IEEE Trans. Image Process. 22 (7) (2013) 2559–2573, http://dx.doi.org/10.1109/TIP.2013.2253477.
[28] M. Pantic, L.J. Rothkrantz, Expert system for automatic analysis of facial expressions, Image Vis. Comput. 18 (11) (2000) 881–905.
[29] C.P. de Campos, Q. Ji, Efficient structure learning of Bayesian networks using constraints, J. Mach. Learn. Res. 12 (3) (2011) 663–689.
[30] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1988.
[31] Q. Ji, RPI Intelligent Systems Lab (ISL) Image Databases ⟨http://www.ecse.rpi.edu/homepages/cvrl/database/database.html⟩.
[32] Y. Tong, Y. Wang, Z. Zhu, Q. Ji, Robust facial feature tracking under varying face pose and facial expression, Pattern Recognit. 40 (11) (2007) 3195–3208.
[33] M.S. Sorower, A Literature Survey on Algorithms for Multi-Label Learning, Technical Report, Oregon State University, Corvallis, OR, USA, December 2010.
[34] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, Bp4d-Spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) 692–706.
[35] P. Ekman, W. V. Friesen, FACS – Facial Action Coding System, 1978 ⟨http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm⟩.
[36] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699.
[37] W.-S. Chu, F. Torre, J. Cohn, Selective transfer machine for personalized facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3515–3522.
[38] X. Zhang, M.H. Mahoor, S.M. Mavadati, J.F. Cohn, A l p-norm MTMKL framework for simultaneous detection of multiple facial action units, in: 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE Piscataway, NJ, 2014, pp. 1104–1111.
[39] T. Baltrusaitis, M. Mahmoud, P. Robinson, Cross-dataset learning and person-specific normalisation for automatic action unit detection, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, IEEE Piscataway, NJ, 2015, pp. 1–6.
[40] J. Nicolle, K. Bailly, M. Chetouani, Facial action unit intensity prediction via hard multi-task metric learning for kernel regression, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6,IEEE Piscataway, NJ, 2015, pp. 1–6.
[41] A. Gudi, H.E. Tasli, T.M. den Uyl, A. Maroulis, Deep learning based FACS action unit occurrence and intensity estimation, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, IEEE Piscataway, NJ, 2015, pp. 1–5.
[42] A. Yuce, H. Gao, J.-P. Thiran, Discriminant multi-label manifold embedding for facial action unit detection, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, IEEE Piscataway, NJ, 2015, pp. 1–6.

Shangfei Wang received her BS in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996. She received her MS in circuits and systems and the Ph.D. in signal and information processing from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 1999 and 2002 respectively. From 2004 to 2005, she was a postdoctoral research fellow in Kyushu University, Japan. Between 2011 and 2012, Dr. Wang was a visiting scholar at Rensselaer Polytechnic Institute in Troy, NY, USA. She is currently an associate professor of the School of Computer Science and Technology, USTC. Dr. Wang is an IEEE and ACM member. Her research interests cover computation intelligence, affective computing, and probabilistic graphical models. She has authored or co-authored over 70 publications. Dr. Wang is a senior member of the IEEE and a member of the ACM.

Quan Gan received his BS in computer science from Hefei University of Technology in 2013, and he is currently pursuing his MS in computer science from the University of Science and Technology of China, Hefei, China. His research interest is affective computing.

Qiang Ji received his Ph.D. 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 Robotics Institute at Carnegie Mellon University, the Department of Computer Science at the 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 by major companies including Honda and Boeing. Prof. Ji is an editor of several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. Prof. Ji is a fellow of the IEEE and IAPR.