

Capturing Feature and Label Relations Simultaneously for Multiple Facial Action Unit Recognition

Shangfei Wang¹, Senior Member, IEEE, Shan Wu, Guozhu Peng¹, and Qiang Ji², Fellow, IEEE

Abstract—Although both feature dependencies and label dependencies are crucial for facial action unit (AU) recognition, little work addresses them simultaneously till now. In this paper, we propose a 4-layer Restricted Boltzmann Machine (RBM) to simultaneously capture feature level and label level dependencies to recognize multiple AUs. The middle hidden layer of the 4-layer RBM model captures dependencies among image features for multiple AUs, while the top latent units capture the high-order semantic dependencies among AU labels. Furthermore, we extend the proposed 4-layer RBM for facial expression-augmented AU recognition, since AU relations are influenced by expressions. By introducing facial expression nodes in the middle visible layer, facial expressions, which are only required during training, facilitate the estimation of both feature dependencies and label dependencies among AUs. Efficient learning and inference algorithms for the extended model are also developed. Experimental results on three benchmark databases, i.e., the CK+ database, the DISFA database and the SEMAINE database, demonstrate that the proposed approaches can successfully capture complex AU relationships from features and labels jointly, and the expression labels available only during training are benefit for AU recognition during testing for both posed and spontaneous facial expressions.

Index Terms—AU recognition, expression-augmented, RBM

1 INTRODUCTION

Facial expression plays an important role in information transmission during people's communication. Therefore, automatic face expression recognition has attracted increasing attention in the fields of both human behavior research [1] and computer vision [2], [3], [4], [5], [6]. There are two ways to describe expressions: multiple expression categories and multiple Facial Action Units (AUs) [7]. Expression categories depict the facial behavior in a global aspect, while AUs represent facial muscle actions locally. Numerous algorithms are proposed to deal with these two kinds of expression recognition problems during the past several years [5], [6], [8], [9], [10], [11], [12]. In this paper, we tackle the problem of facial action unit recognition.

Current research of facial action unit recognition mainly recognizes each action unit independently or detects a fixed number of AU combinations. They either ignore AU relations or fail to handle hundreds of variations in AU combinations. Behavior research indicates that there exist co-

current and mutual exclusive relations among AUs, and multiple AUs can lead to a huge variety of complex facial behaviors with different combinations. For example, five facial action units, i.e., AU4, AU6, AU9, AU15, and AU17, appear simultaneously on the face of Fig. 1a, and both AU15 and AU17 present in Fig. 1b. AU15 (lip corner depressor) and AU12 (lip corner puller) can not appear together due to the structure relations of facial muscles. Such AU relations can be leveraged for better AU recognition. Furthermore, these inherent relations exist not only in AU labels but also facial appearance. For instance, as illustrated in Fig. 1a, when AU6 and AU9 occur together, wrinkle appears between eyebrows along with eyes narrowing; when AU6 combines with AU12 depicted in Fig. 1b, lip corners are pulled and eyes are narrowed with no wrinkles between eyebrows.

AU relationships are context-dependent. They are influenced by a lot of related factors such as the facial expressions. Almost all facial expressions consist of several certain combinations of AUs. For example, as shown in Figs. 1d and 1e, happiness always happens with AU6, AU12 and AU25, and surprise usually appears with AU1 and AU25. Different expressions may generate the same AUs. For example, AU25 could appear on both a happy face and a surprise face. When AU25 appears on a happy face, the cheeks raise dramatically; when AU25 appears on a surprise face, the cheeks do not raise. They cause different facial appearances on the cheek and mouth regions. Such inherent relations between expressions and among AUs can facilitate AU recognition.

Only recently, several works exploit AU dependencies to improve AU recognition. Tong et al. [10], Wang et al. [11]

- S. Wang, S. Wu, and G. Peng are with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230000, China.
E-mail: sfwang@ustc.edu.cn, {sa14ws, gzpeng}@mail.ustc.edu.cn.
- Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180.
E-mail: qji@ecse.rpi.edu.

Manuscript received 13 Apr. 2017; revised 23 June 2017; accepted 1 Aug. 2017. Date of publication 8 Aug. 2017; date of current version 12 Sept. 2019.

(Corresponding author: Shangfei Wang.)

Recommended for acceptance by S. P. Zafeiriou.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TAFFC.2017.2737540

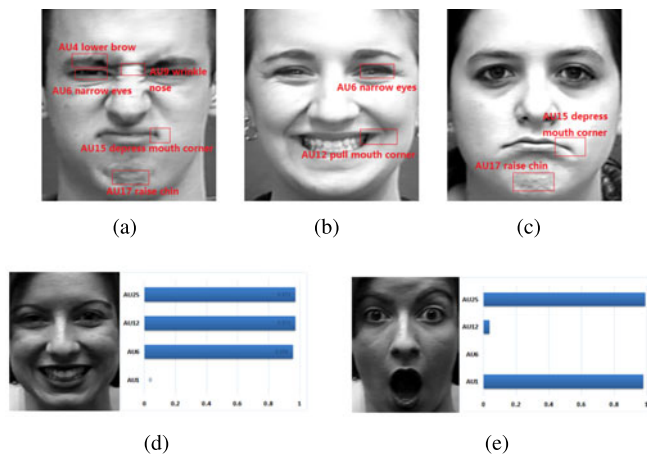


Fig. 1. Samples of facial action units (a) lower brow(AU4), narrow eyes (AU6), wrinkle nose(AU9), depress mouth corner(AU15), raise chin (AU17); (b) narrow eyes(AU6), pull mouth corner(AU12); (c) depress mouth corner(AU15), raise chin(AU17); (d) the probability of AUs appeared for happy; (e) the probability of AUs appeared for surprise.

and Song et al. [12] modeled AU dependencies among AU labels, without considering AU dependencies inherited in facial appearance. Zhu et al. [13] captured AU relations in both AU labels and facial features separately, not jointly. Zhang et al. [14] and Zhao et al. [15] integrated fixed local AU relations existing in AU labels into AU classifiers. Eleftheriadis et al. [16] combined co-occurrent AU dependencies into latent space and classifier learning through a multi-conditional latent variable model.

Compared to AU-relations modeling for AU recognition, capturing AU-expression dependencies for AU recognition has attracted even less attention. To the best of our knowledge, only three works [11], [17], [18] recognize AUs assisted by expressions. All of them modeled AU-expression dependencies among AU labels, without considering AU-expression dependencies inherited in facial appearance.

Therefore, in this paper, we first propose a novel hierarchical model to recognize multiple facial action units by exploiting AU relations from both feature level and label level jointly. Specifically, we propose a 4-layer Restricted Boltzmann Machine (FRBM), which includes two hidden layers in addition to one feature layer and one ground-truth AU label layer. The top hidden layer models the high-order dependencies among the ground-truth AUs through the connections to all AU labels. By connecting features and AU labels, the bottom hidden layer not only captures the dependencies among features, but also represents the salient information for the input features. The connections of four layers are jointly trained with interactions among layers for learning shared features and AU relations simultaneously.

Furthermore, we extend the proposed FRBM model by incorporating facial expression to enhance AU recognition. We refer to the extended model as FRBM+. By adding facial expression nodes in the middle visible layer, i.e., ground-truth AU label layer, the top hidden layer models the high-order dependencies among the ground-truth expression and AU labels through the connections to all AU labels and expression label. The expression labels are only required during training. By adding expression nodes, the extended model, FRBM+, does not dramatically increase the complexity of the proposed FRBM model, but can successfully

capture extra expression-dependent AU relations for AU recognition. We further propose an efficient learning and inference algorithm for this model.

AU recognition experiments demonstrate the effectiveness of proposed method for jointly exploiting AU dependencies in features and labels and their superior performances of AU recognition to existing methods. Expression-augmented AU recognition experiments further demonstrate that the inherent relations between AUs and expressions are crucial for AU recognition.

2 RELATED WORK

2.1 AU Recognition by Exploiting AU Relations

The main stream of facial action unit analysis research either recognizes each action unit independently or recognizes fixed action unit combinations. The former ignores the inherent dependencies among AUs, and the latter only handles several fixed AU combinations, and can not detect thousands of possible combinations. Only recently, a few researchers began to exploit AU relations for AU recognition.

Tong et al. [10] proposed a Dynamic Bayesian Network (DBN) to capture probabilistic relations among AUs and temporal changes in facial action development. First, Gabor features and Support Vector Machine (SVM) are used to recognize each AU. Then the recognized AUs are used as evidence to the DBN for inferring various AUs. Tong et al.'s work successfully modeled AU relations from target labels, and demonstrated the benefit of AU relations for AU recognition. However, due to the Markov assumption, their proposed DBN is limited to capturing the local relationships between pairs of AUs, such as co-occurrence, co-absence and mutual exclusion. To overcome this, Wang et al. [11] proposed a three-layer Restricted Boltzmann Machine to capture global relations among all AUs, and integrate the AU measurements with the high-level AU semantical relationships for AU recognition. More recently, Song et al. [12] modeled AU sparsity and co-occurrence using a Bayesian compressed sensing model. They employed a AU-conditional thresholding to obtain the AU groups which may capture the global dependencies among AUs to some extent. These works successfully model AU relations from target labels, but ignore AU inherent relations in image features, which are crucial for AU analysis.

Zhu et al. [13] proposed multiple facial action units recognition by modeling their relations from both features and target labels. First, a multi-task feature learning method is adopted to learn the shared features for each AU group. Second, a Bayesian Network (BN) is used to model relations among action units from the target labels. After that, the learned Bayesian network employs the recognition results of the multi-task learning, and realizes multiple facial action units recognition by probabilistic inference. The multi-task feature learning and the Bayesian network are learned separately. Such independent modeling could produce inconsistent dependencies between feature-level and label-level.

Zhang et al. [14] utilized multi-task multiple kernel learning to detect multiple AUs simultaneously. They 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 group share the same kernel. A multi-task multiple kernel learning is used to learn AU classifiers simultaneously.

Zhao et al. [15] introduced joint-patch and multi-label learning to model dependencies among both features and AUs. Specifically, they select a sparse subset of facial patches and learn multiple AU classifiers simultaneously under the constraints of group sparsity and local AU relations (i.e., positive correlation and negative competition). These two works integrate AU relations existing in AU labels into AU classifiers by using multi-task multiple kernel learning as well as joint-patch and multi-label learning respectively. Therefore, they model AU relations in both labels and features jointly. However, they only capture fixed local dependencies.

Instead of integrating predefined local label dependencies into classifiers in the original feature space or kernel space as the above works, Eleftheriadis et al. [16] proposed a multi-conditional latent variable model to combine global label dependencies into latent space and classifier learning. Specifically, the image features are projected onto a shared manifold. Then, multiple AU classifiers are learned simultaneously on the manifold. The subspace is regularized by constraints, which encode local and global co-occurrence dependencies among AU labels. The label relations exploited in this work are limited to co-occurrence, without considering mutual exclusion relationships among AU labels.

Current work of AU recognition through exploiting AU relations fails to learn global probabilistic inherent dependencies (i.e., co-existence and mutual exclusion) in both facial features and AU labels jointly. Therefore, in this paper, we propose a novel hierarchical model to jointly capture relations in these two levels. Our model consists of two latent layers. The top one aims to model high-order dependencies among AUs, and the bottom one is designed to link the AUs and features so as to capture the commonalities and characteristics among different AUs. The learning processes in feature level and label level are dependent on each other, in this case, the relationships in these two levels are compatible. Furthermore, through adding expression nodes in the middle visible layer, our model can be extended to jointly capture feature dependencies and label dependencies with the help of expressions, which are only required during training.

2.2 AU Recognition Augmented by Expressions

Compared to AU-relations modeling for AU recognition, capturing AU-expression dependencies for AU recognition has attracted even less attention. To the best of our knowledge, there are so far only three works that recognize AUs assisted by expressions. Wang et al. [17] designed an AU recognition method with the help of expression labels as hidden knowledge under incomplete AU labeling. They proposed to construct a BN model to capture the dependencies not only among AUs but also between AUs and expressions. During training, the image features and expression labels are complete, while the AU labels may be missing. Structure Expectation Maximization (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.

Due to the Markov assumption, the BN model can only handle local dependencies among AUs and expressions. Wang et al. [11] proposed to use a 3-way RBM to capture the global dependencies between expressions and AUs for AU recognition. The expression labels are used as privileged information, which is only required during training. The 3-way RBM model can be regarded as a mixture model, therefore, it models the relations between each expression and AUs independently, ignoring the shared dependencies among multiple expressions and AUs.

Ruiz et al. [18] proposed to learn AU classifiers from unannotated facial images under the help of another large scale of facial images with expression labels, but without AU labels. Their method consists of both AU classifiers from images and expression classifiers from AUs. Prior knowledge about the relation between expressions and AUs is employed to learn expression classifier from AUs first, then AU classifiers from images can be learned through embedding the output of AU classifiers as the input of expression classifier. In the first step, the prior knowledge on statistical correlations between AUs and expressions is used to generate pseudo AU labels from the ground-truth expression labels.

The three works successfully leverage AU-expression dependencies from target labels for AU recognition, but ignore inherent AU-expression relations in image features, which are crucial for AU analysis. Therefore, in this paper, we extend our proposed hierarchical model to jointly capture AU-expression dependencies from both target labels and features. By adding facial expression nodes in the middle visible layer, the top hidden layer models the high-order dependencies among the ground-truth expression and AU labels through the connections to all AU nodes and expression nodes, and the bottom hidden layer captures the dependencies among features for a multi-task classifier to estimate multiple AU labels and expression. The connections of four layers are jointly trained with interactions among layers for learning shared features and label relations simultaneously. The expression labels are only required during training to assist AU recognition.

3 METHODS

We propose two models to perform AU recognition. The first model, FRBM, aims at capturing the high-order AU dependencies in label level and the commonalities and characters among AUs in feature level, which is depicted in Fig. 2a. The second model, FRBM+, is proposed to further capture the global relations among AUs and expressions so as to facilitate the estimation of relations among AUs, as shown in Fig. 2b. Compared Fig. 2a with Fig. 2b, the second layer y in Fig. 2a represents the AU labels, while in Fig. 2b, the second layer consists of two parts, i.e., y and \hat{y} , where y represents AU labels and \hat{y} represents encoded expression label. Therefore, the model in Fig. 2a only captures the high-order dependencies among AU labels, while the model in Fig. 2b exploits not only the dependencies among AUs but also the inherent relations between AUs and expressions.

3.1 AU Recognition Through AU-Relation Modeling

Fig. 2a depicts the structure of our proposed FRBM model for AU recognition by capturing AU dependencies from

features and labels. \mathbf{x} is the features which are continuous variables, $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ are binary hidden layers, and \mathbf{y} represents the states of AUs. In this model, the top hidden layer $\mathbf{h}^{(1)}$ allows to capture the global dependencies among multiple AUs, and the middle hidden layer $\mathbf{h}^{(2)}$ is designed to capture the commonalities and characters among AUs.

The energy function of FRBM is defined in

$$\begin{aligned} E(\mathbf{y}, \mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Theta) &= - \sum_t h_t^{(1)} d_t - \sum_j y_j c_j - \sum_i h_i^{(2)} b_i \\ &+ \frac{1}{2} \sum_k \frac{(x_k - a_k)^2}{\delta_k^2} - \sum_t \sum_j h_t^{(1)} W_{tj}^{(1)} y_j \\ &- \sum_i \sum_j h_i^{(2)} W_{ij}^{(2)} y_j - \sum_i \sum_k h_i^{(2)} W_{ik}^{(3)} \frac{x_k}{\delta_k}, \end{aligned} \quad (1)$$

where $\{d_t\}$, $\{c_j\}$, $\{b_i\}$ and $\{a_k\}$ denote the bias of each node in $\mathbf{h}^{(1)}$, \mathbf{y} , $\mathbf{h}^{(2)}$ and \mathbf{x} respectively, and $\Theta = \{\mathbf{d}, \mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}\}$ represents the parameters. δ_k is the standard deviation of the Gaussian noise for x_k . Generally, learning the variance of the noise for each unit is difficult. But it is easier to normalise each component of the data to have zero mean and unit variance and then to use noise free reconstructions. So for convenience, we normalize the data to make δ_k equal 1.

To obtain the joint distribution of input data and labels, we need to marginalize over all the hidden nodes including $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ according to

$$p(\mathbf{y}, \mathbf{x}; \Theta) = \frac{\sum_{\mathbf{h}^{(1)}} \sum_{\mathbf{h}^{(2)}} \exp(-E(\mathbf{y}, \mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Theta))}{Z(\Theta)}, \quad (2)$$

where $Z(\Theta) = \sum_{\mathbf{y}, \mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}} \exp(-E(\mathbf{y}, \mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Theta))$ is the partition function.

During training phase, a Maximum Likelihood Estimation (MLE) method is adopted to learn the parameters according to Equation (3). The gradient of each parameter is defined in Equation (4). Due to the complexity to calculate $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \mathbf{x}; \Theta)$ and $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{y}, \mathbf{x}; \Theta)$, we employ the Contrastive Divergence (CD) algorithm [19]

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\Theta), \quad \mathcal{L}(\Theta) = \log p(\mathbf{y}^{(i)}, \mathbf{x}^{(i)}; \Theta) \quad (3)$$

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \mathbf{x}; \Theta)} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{y}, \mathbf{x}; \Theta)}. \quad (4)$$

During inference, we recognize each facial image with the most probable states of AUs according to

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}} p(\mathbf{y} | \mathbf{x}; \Theta). \quad (5)$$

The probability of each state can be obtained by calculating the posterior probability, defined in

$$p(\mathbf{y} | \mathbf{x}; \Theta) = \frac{\exp(\mathbf{c}^T \mathbf{y}) \prod_t (1 + \exp(\alpha_{yt})) \prod_i (1 + \exp(\beta_{yi}))}{\sum_{\mathbf{y}^*} \exp(\mathbf{c}^T \mathbf{y}^*) \prod_t (1 + \exp(\alpha_{y^*t})) \prod_i (1 + \exp(\beta_{y^*i}))}, \quad (6)$$

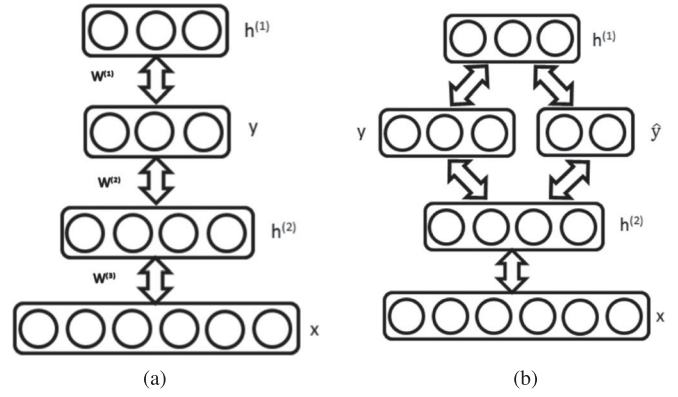


Fig. 2. (a) *FRBM*: The proposed method for AU recognition by capturing high-order relations among multiple AUs. The second layer \mathbf{y} represents the AU labels; (b) *FRBM+*: The proposed method for AU recognition enhanced by expression by capturing the relations among AUs and expression. The second layer consists of two parts: AU labels \mathbf{y} and encoded expression label $\hat{\mathbf{y}}$.

where $\alpha_{yt} = d_t + \sum_j W_{tj}^{(1)} y_j$, $\beta_{yi} = b_i + \sum_j W_{ij}^{(2)} y_j + \sum_k W_{ik}^{(3)} x_k$. From this equation, it is easy to find that the computation cost is exponential with the number of labels, so we adopt a Gibbs Sampling method to infer $p(\mathbf{y} | \mathbf{x})$.

3.2 AU Recognition Enhanced by Expressions

As discussed in Section 1, the semantic relations between AUs and expressions are crucial for AU recognition. Hence in this section, we revise FRBM to incorporate expression information to enhance AU recognition, which is shown in Fig. 2b, i.e., *FRBM+*. Similarly, \mathbf{x} represents the features, $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ are binary hidden layers, and $\mathbf{h}^{(2)}$ is used to capture the commonalities and characters among AUs. Unlike *FRBM*, the second layer in *FRBM+* consists of two parts: AUs and expression. \mathbf{y} represents the AU labels, and $\hat{\mathbf{y}}$ stands for expression label. To match the form of AUs, the facial expressions are presented in their binary form. Specifically, for K expressions, $n = \lceil \log K \rceil$ latent units are needed to encode all expressions. Therefore, in this model, $\mathbf{h}^{(1)}$ can not only model the global dependencies among AUs, but also capture the global relations among AUs and expressions simultaneously.

The energy function of *FRBM+* is defined in Equation (7). $\Theta = \{\mathbf{d}, \mathbf{c}, \mathbf{b}, \mathbf{a}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}\}$ represents the parameters. $\{d_t\}$, $\{c_j\}$, $\{c_s\}$, $\{b_i\}$ and $\{a_k\}$ denote the bias of each node in $\mathbf{h}^{(1)}$, \mathbf{y} , $\hat{\mathbf{y}}$, $\mathbf{h}^{(2)}$ and \mathbf{x} respectively. δ_k is the standard deviation of the Gaussian noise for x_k . Similar to *FRBM*, we normalize the data to make δ_k equal 1

$$\begin{aligned} E(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \Theta) &= - \sum_t h_t^{(1)} d_t - \sum_j y_j c_j - \sum_s \hat{y}_s c_s - \sum_i h_i^{(2)} b_i \\ &+ \frac{1}{2} \sum_k \frac{(x_k - a_k)^2}{\delta_k^2} - \sum_t \sum_j h_t^{(1)} W_{tj}^{(1)} y_j - \sum_t \sum_s h_t^{(1)} W_{ts}^{(1)} \hat{y}_s \\ &- \sum_i \sum_j h_i^{(2)} W_{ij}^{(2)} y_j - \sum_i \sum_s h_i^{(2)} W_{is}^{(2)} \hat{y}_s - \sum_i \sum_k h_i^{(2)} W_{ik}^{(3)} \frac{x_k}{\delta_k}. \end{aligned} \quad (7)$$

Comparing Equation (7) with Equation (1), we can find that the proposed AU recognition model *FRBM* is a special case of the proposed expression-assisted AU recognition model *FRBM+* in the case of $\hat{\mathbf{y}} = 0$.

3.3 Model Learning

We employ the MLE method to learn the parameters of the model according to Equation (8). The gradient of each parameter $\theta \in \Theta$ is defined in Equation (9). A stochastic gradient ascent method is used to update the parameters

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\Theta), \mathcal{L}(\Theta) = \log p(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)} | \mathbf{x}^{(i)}, \Theta) \quad (8)$$

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \hat{\mathbf{y}} | \mathbf{x}, \Theta)} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{y}, \hat{\mathbf{y}}, \mathbf{x}, \Theta)} \quad (9)$$

Calculating the gradient involves inferring $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \hat{\mathbf{y}} | \mathbf{x}, \Theta)$ and $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{y}, \hat{\mathbf{y}}, \mathbf{x}, \Theta)$. These two conditional probabilities are intractable to compute. So we adopt a CD algorithm to learn the parameters. Specifically, $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \hat{\mathbf{y}} | \mathbf{x}, \Theta)$ can be estimated by sampling $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{y}, \hat{\mathbf{y}}$ according to Equations (10), (11), (12), and (13) respectively, where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. Similarly, we can estimate $p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{y}, \hat{\mathbf{y}}, \mathbf{x}, \Theta)$ by sampling $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ according to Equations (10) and (11) separately. Detailed learning algorithm is described in Algorithm 1

$$p(h_i^{(1)} = 1 | \mathbf{y}, \hat{\mathbf{y}}) = \sigma(d_i + \sum_j W_{tj}^{(1)} y_j + \sum_s W_{ts}^{(1)} \hat{y}_s) \quad (10)$$

$$\begin{aligned} p(h_i^{(2)} = 1 | \mathbf{y}, \hat{\mathbf{y}}, \mathbf{x}) \\ = \sigma(b_i + \sum_j W_{ij}^{(2)} y_j + \sum_s W_{is}^{(2)} \hat{y}_s + \sum_k W_{ik}^{(3)} x_k) \end{aligned} \quad (11)$$

$$p(y_j = 1 | \mathbf{h}^{(1)}, \mathbf{h}^{(2)}) = \sigma(c_j + \sum_t h_t^{(1)} W_{tj}^{(1)} + \sum_i h_i^{(2)} W_{ij}^{(2)}) \quad (12)$$

$$p(\hat{y}_s = 1 | \mathbf{h}^{(1)}, \mathbf{h}^{(2)}) = \sigma(c_s + \sum_t h_t^{(1)} W_{ts}^{(1)} + \sum_i h_i^{(2)} W_{is}^{(2)}) \quad (13)$$

Algorithm 1. Parameters Learning Algorithm of FRBM+

Input: Training data $D = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$.

Output: $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$.

```

1: repeat
2:   for each  $i \in [1, N]$  do
3:     % Positive phase
4:     Initialize  $\mathbf{y}^0, \mathbf{x}^0$  with  $\mathbf{y}^{(i)}, \mathbf{x}^{(i)}$ 
5:     Calculate  $\hat{\mathbf{h}}^{(1)0}$  by Equation (10)
6:     Calculate  $\hat{\mathbf{h}}^{(2)0}$  by Equation (11)
7:
8:     % Negative phase
9:     Sample  $\mathbf{h}^{(1)0}$  from  $\hat{\mathbf{h}}^{(1)0}$ 
10:    Sample  $\mathbf{h}^{(2)0}$  from  $\hat{\mathbf{h}}^{(2)0}$ 
11:    Calculate  $\mathbf{y}^1, \hat{\mathbf{y}}^1$  by Equations (12) and (13)
12:    Calculate  $\hat{\mathbf{h}}^{(1)1}$  by Equation (10)
13:    Calculate  $\hat{\mathbf{h}}^{(2)1}$  by Equation (11)
14:
15:    % Update phase
16:    for  $\theta \in \Theta$  do
17:      Calculate the gradient of  $\theta$  by Equation (9)
18:      Update  $\theta$  by the gradient of  $\theta$ 
19:    end for
20:  end for
21: until Converges

```

3.4 Inference

Given a query sample \mathbf{x} , we infer the state of AUs according to Equation (14). And the posterior probability $p(\mathbf{y}, \hat{\mathbf{y}} | \mathbf{x}, \Theta)$ is defined in Equation (15),

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{arg max}} p(\mathbf{y}, \hat{\mathbf{y}} | \mathbf{x}, \Theta) \quad (14)$$

$$p(\mathbf{Y} | \mathbf{x}; \Theta) = \frac{\exp(\mathbf{c}^T \mathbf{Y}) \prod_t (1 + \exp(\alpha_{Yt})) \prod_i (1 + \exp(\beta_{Yi}))}{\sum_{\mathbf{Y}^*} \exp(\mathbf{c}^T \mathbf{Y}^*) \prod_t (1 + \exp(\alpha_{Y^*t})) \prod_i (1 + \exp(\beta_{Y^*i}))} \quad (15)$$

where $\mathbf{Y} = (\mathbf{y}, \hat{\mathbf{y}})$, $\alpha_{Yt} = d_t + \sum_j W_{tj}^{(1)} y_j + \sum_s W_{ts}^{(1)} \hat{y}_s$, $\beta_{Yi} = b_i + \sum_j W_{ij}^{(2)} y_j + \sum_s W_{is}^{(2)} \hat{y}_s + \sum_k W_{ik}^{(3)} x_k$. Since the computational cost is exponential with the number of labels, we adopt the Gibbs Sampling method to infer $p(\mathbf{y}, \hat{\mathbf{y}} | \mathbf{x})$ for large number of AUs. Algorithm 2 lists two methods for inference, where C is the number of sampling, and M is the steps of Gibbs Sampling. We use a validation set to determine the value of C and M .

Algorithm 2. Inference of FRBM+ by Gibbs Sampling

Input: query sample \mathbf{x} , $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$.

Output: the predicted states of AUs \mathbf{y}^* for \mathbf{x}

```

1: method 1: Infer the states of AUs by Equation (14).
2: method 2: Use Gibbs Sampling to infer  $\mathbf{y}^*$ 
3: repeat
4:   for chain = 1  $\rightarrow$  C do do
5:     randomly initialize  $\mathbf{y}^0$ 
6:     for  $r = 0 \rightarrow M$  do do
7:       Sample  $\mathbf{h}^{(1)r} \sim p(\mathbf{h}^{(1)} | \mathbf{y}, \hat{\mathbf{y}})$  with Equation (10)
8:       Sample  $\mathbf{h}^{(2)r} \sim p(\mathbf{h}^{(2)} | \mathbf{y}, \hat{\mathbf{y}}, \mathbf{x})$  with Equation (11)
9:       Sample  $\mathbf{y}^{r+1} \sim p(\mathbf{y} | \mathbf{h}^{(1)}, \mathbf{h}^{(2)})$  with Equation (12)
10:      Sample  $\hat{\mathbf{y}}^{r+1} \sim p(\hat{\mathbf{y}} | \mathbf{h}^{(1)}, \mathbf{h}^{(2)})$  with Equation (13)
11:    end for
12:  end for
13:  for  $j = 1 \rightarrow n$  do
14:    collect the last K samples of  $(\mathbf{y}_j, \hat{\mathbf{y}}_j)$  from each chain
15:    calculate  $p(\mathbf{y}_j, \hat{\mathbf{y}}_j | \mathbf{x})$  based on the collected samples
16:  end for
17: until Converges

```

4 EXPERIMENTS

4.1 Experimental Conditions

In our experiments, three benchmark databases are used: the extended Cohn-Kanade (CK+) database [20], the Denver Intensity of Spontaneous Facial Actions (DISFA) [21], and the the SEMAINE database [22].

The CK+ database contains 593 posed expression image sequences, starting from neutral frame and ending at the peak frame. The last frame for each image sequence is FACS coded. And it also provides seven expression categories, i.e., Angry, Contempt, Disgust, Fear, Happy, Sadness and Surprise, for 327 samples. Similar to [17], 13 AUs whose frequencies are more than 10 percent are selected, i.e., AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU17, AU23, AU24, AU25, and AU27. Let $\{x_i, y_i\}$ and $\{x_{0i}, y_{0i}\}$ represent the facial point of peak frame and the corresponding neutral frame respectively. Similar to [17], the differences of facial

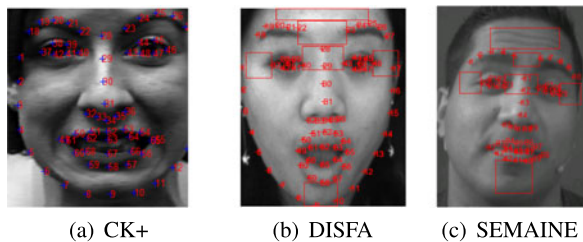


Fig. 3. (a) Facial points on the CK+ database; (b)(c) Facial points and the regions used to extract gabor features on the DISFA database and SEMAINE database.

points, denoted as $[x_1 - x_{01}, y_1 - y_{01}, \dots, x_c - x_{0c}, y_c - y_{0c}]$, are used as features. $c = 68$ is the number of facial points. Fig. 3a depicts the details of the 68 facial points. Furthermore, before feature extraction, the face images are aligned according to the coordinates of the eyeball.

The DISFA database contains 27 spontaneous facial expression videos of 27 subjects while watching YouTube videos. Each frame is rated in terms of the AU intensity on a six-point scale. In our work, we treat each AU with intensity larger than zero as active. Similar to [16], we select seven highly correlated AUs, i.e., AU1, AU2, AU4, AU6, AU12, AU15 and AU17, to validate our methods. Frames with at least 3 active AUs are selected in our work, resulting in 5,863 samples. Since the DISFA database does not explicitly indicate the neutral frame and peak frame of the videos, both appearance and geometric features are adopted in our work, which are similar to [13]. For the geometric features, the absolute coordinates are used. For the appearance features, the Gabor features are extracted from the regions of the forehead, between the eyebrows, between the eyes, outer corner of eyes, and lower jaw, as shown in Fig. 3b.

The samples in the SEMAINE database are naturally induced by operators during the conversation. Therefore the dataset contains speech related mouth and face movements, and significant amounts of both in- and outof-plane head rotations. All these make the recognition task much more challenging. So far a total of 180 frames from 8 sessions of two subjects on the SEMAINE database are FACS coded with experts. Similar to [11], we recognize 10 AUs, i.e., AU1, AU2, AU4, AU5, AU6, AU7, AU12, AU17, AU25, AU26, which are present for at least 15 times. We extract the same features as those of the DISFA database, as shown in Fig. 3c. Table 1 shows the data distributions of three databases.

To validate the effectiveness of our proposed models, we conduct two experiments: AU recognition through AU-relation modeling, i.e., FRBM, and AU recognition with expression assisting, i.e., FRBM+. For the first experiment, all databases are used; for the second experiment, only the CK+ database and the SEMAINE database are adopted, since the

TABLE 1
Data Distribution on Three Databases

	AU	1	2	4	5	6	7	9
CK+	number	165	117	170	100	113	119	60
	AU	12	17	23	24	25	27	
	number	83	133	55	49	218	73	
DISFA	AU	1	2	4	6	12	15	17
	number	2,669	1,737	4,683	2,725	1,789	2,438	3,191
	AU	1	2	4	5	6	7	12
SEMAINE	number	51	52	34	16	54	41	64
	AU	17	25	26				
	number	28	107	70				

DISFA database does not provide expression labels. On the CK+ database and the DISFA database, a leave-one-subject-out cross validation is adopted. Since the FACS coded samples in the SEMAINE database are from two subjects, we employ the 10-fold cross validation on this database.

4.2 Experimental Results of AU Recognition Through AU-Relation Modeling

4.2.1 Analyses of AU Recognition

To validate the effectiveness of our method, we compare our approach with the related works mentioned in Section 2. Due to the availability of the codes for BGCS [12] and HRBM [11], we re-conduct the experiments of BGCS and HRBM using our data. Eleftheriadis et al. [16] have already compared their method with many related works, such as l_p -MTMKL [23], on the CK+ database and the DISFA database. So we directly compare our results on these two databases with those in [16]. Since the features and samples used in our work are not exactly the same as [16], the comparisons with MC-LVM and l_p -MTMKL are only for reference. The results are listed in Tables 2 and 3. Due to the unavailability of the codes for MC-LVM [16] and l_p -MTMKL [23], we can not compare our results with theirs on the SEMAINE database. Table 4 lists the results on the SEMAINE database.

From the results in Tables 2, 3, and 4, we find that our method performs best among all the state-of-the-art approaches on three databases.

On the CK+ database, our method achieves the highest F1-score. Considering the AUs used in [16], the average F1-score of our approach is 0.8267, which is 3 and 9 percent higher than MC-LVM and l_p -MTMKL respectively. Compared with BGCS, the average F1-score of ours is nearly 14 percent higher. Furthermore, our method performs better on all AUs. And the per-AU F1-score of ours on 13 AUs are higher than 70 percent except for AU23 and AU24, since they are hard to detect and the frequencies of occurrence for these two AUs are fewer than others (shown in Table 1).

TABLE 2
AU Recognition Results Through AU-Relation Modeling on the CK+ Database

AU	1	2	4	5	6	7	9	12	17	23	24	25	27	Avg.
FRBM	0.8978	0.8861	0.8176	0.7980	0.7907	0.7018	0.8870	0.8471	0.8456	0.6792	0.5918	0.9557	0.9007	0.8153
BGCS [12]	0.7411	0.8263	0.6426	0.6533	0.6411	0.6023	0.7869	0.7711	0.6531	0.3188	0.4130	0.8740	0.8718	0.6766
HRBM [11]	0.8896	0.9231	0.8121	0.7729	0.7623	0.7130	0.8667	0.7978	0.8358	0.6604	0.5435	0.8949	0.8553	0.7944
MC-LVM [16]	0.8439	0.8655	0.8160	-	0.6842	0.6167	-	0.8848	0.8740	-	-	-	-	0.7979
l_p -MTMKL [16]	0.8750	0.8550	0.5143	-	0.7265	0.5882	-	0.8595	0.7544	-	-	-	-	0.7390

TABLE 3
AU Recognition Results Through AU-Relation Modeling on the DISFA Database

AU	1	2	4	6	12	15	17	average
FRBM	0.6194	0.7098	0.8062	0.5215	0.5675	0.5921	0.7531	0.6528
BGCS [12]	0.4998	0.4942	0.9075	0.5604	0.5310	0.5667	0.7360	0.6137
HRBM [11]	0.6256	0.4571	0.4884	0.6346	0.6396	0.5874	0.6165	0.5785
MC-LVM [16]	0.5855	0.6299	0.7285	0.5232	0.8474	0.4944	0.4863	0.6136
l_p -MTMKL [16]	0.4221	0.4581	0.4718	0.6279	0.7633	0.3447	0.4140	0.5003

TABLE 4
AU Recognition Results Through AU-Relation Modeling on the SEMAINE Database

AU	1	2	4	5	6	7	12	17	25	26	average
FRBM	0.800	0.752	0.667	0.400	0.563	0.506	0.683	0.364	0.826	0.676	0.624
BGCS [12]	0.755	0.685	0.623	0.432	0.513	0.380	0.719	0.414	0.726	0.601	0.585
HRBM [11]	0.720	0.686	0.592	0.519	0.520	0.273	0.662	0.360	0.576	0.346	0.525

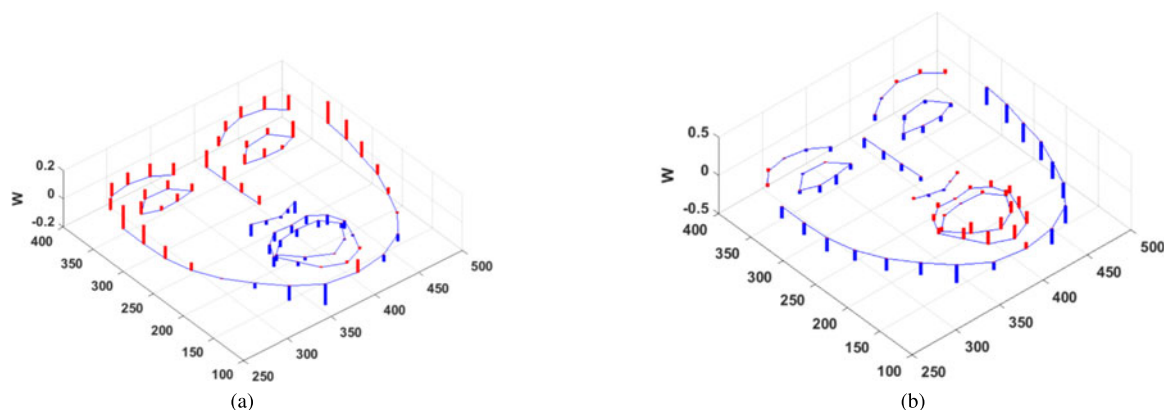


Fig. 4. Semantic relationships in feature level captured by two different hidden nodes of FRBM model trained on the CK+ database. Z-axis: The weight between x and $h^{(2)}$. The red bars represent weight value larger than zero, and the blue bars represent weight value less than zero.

Our method achieves 67 and 59 percent on AU23 and AU24, while the results of BGCS are only 31 and 41 percent respectively. When compared with HRBM, our method outperforms on 11 AUs. And on average, our result is 2.1 percent higher. As for other two works, our approach performs better than l_p -MTMKL on all seven AUs and outperforms MC-LVM on five of seven AUs. These strongly demonstrate the effectiveness of our method.

On the DISFA database, our method performs better than other approaches as well. Compared with BGCS and MC-LVM, our method improves the average F1-score for 4 percent. As for other two methods, i.e., HRBM and l_p -MTMKL, our method improves the performance for 8 and 15 percent separately. When considering per-AU recognition performance, our method is better than the state-of-the-art approaches in most cases. These illustrate that our method also works well on spontaneous facial expressions.

On the SEMAINE database, since both [11] and [16] do not provide the results on this database, we re-conduct the experiment of HRBM [11] using the available code based on our data. From Table 4, we find that our method of FRBM also performs better than HRBM and BGCS on the SEMAINE database. In detail, our method performs better than HRBM on 9 AUs and performs better than BGCS on 7 AUs. And the average F1 Score of FRBM is about 4 and 10 percent higher than those of BGCS and HRBM respectively. This

further demonstrates the effectiveness of our method in dealing with spontaneous data.

Other than modeling dependencies among AUs from labels only [11], [12], intergrading fixed AU label relations with AU classifier learning [23], or capturing co-occurrence of AU labels and dependencies among features [16], our method successfully learns global probabilistic inherent dependencies (i.e., co-existence and mutual exclusion) in both facial features and AU labels jointly. It leads to superior performance of AU recognition.

4.2.2 Analyses of AU Dependencies

As discussed in Section 3, each latent unit can capture a specific pattern, which is measured by the weight between the latent units and features or labels. Specifically, in Fig. 2a, $W^{(1)}$ models the global dependencies among labels in label level, and $W^{(3)}$ encodes the feature relations in feature level. Larger weight indicates high probability of occurrence, while smaller weight denotes high probability of absence.

Figs. 4a and 4b are two instances of feature dependencies captured by two latent units on the CK+ database. The red bars in these figures represent weight value larger than zero, and the blue bars denote weight value less than zero. And larger weight represents a high probability of presence, and smaller weight represents a high probability of absence. Fig. 4a illustrates that there exists a pattern that facial action

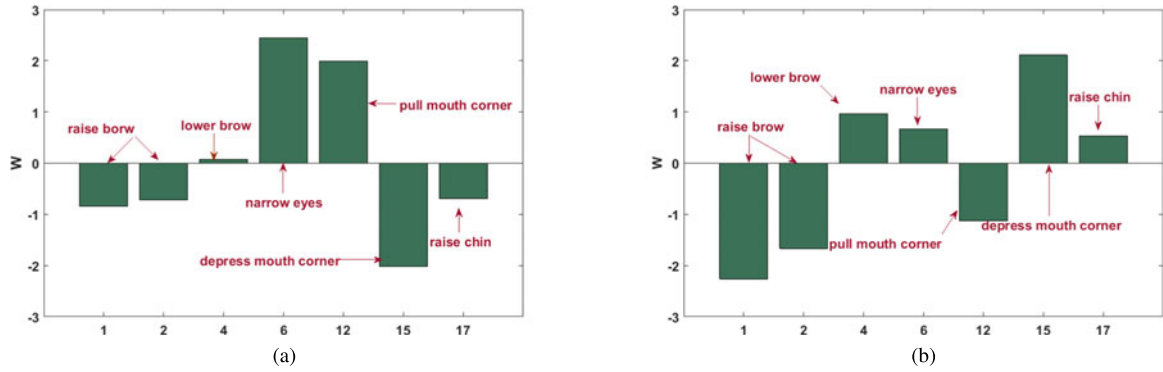


Fig. 5. Semantic relationships in label level captured by two different latent units of FRBM on the DISFA database. X-axis: AU index. Y-axis: The weight between $h^{(1)}$ and y . Larger W indicates high probability of occurrence, while smaller W denotes high probability of absence.

appears more likely around eyes, brows and nose, but not mouth, since most weights on the feature points around eyes, brows and nose are positive, and weights on the feature points of mouth are negative. Fig. 4b shows a completely opposite pattern, which illustrates a pattern that some facial actions are more likely to occur around mouth and jaw due to the positive weights around these regions, but unlikely to occur around eyes, brows and nose.

Figs. 5a and 5b graphically illustrate the global dependencies captured in label level. The relationship depicted in Fig. 5a denotes that someone is likely to narrow his eyes, pull mouth corners and lower brows a little bit, while not raise brows or chin and depress mouth corner. These facial actions may induce a happy expression. Fig. 5b describes a pattern that someone may lower brows, narrow eyes, depress mouth corners and raise chin. These facial actions may present in disgust or angry expression.

4.2.3 Evaluation of Learned Representation

To validate the effectiveness of our model in learning feature representation, we employ t-SNE to depict the embeddings of images represented by the raw image features and the learned representations by the bottom hidden layer of our model, and visualize the effect of individual differences by coloring in terms of subjects [24]. Since the average sample for each subject on the CK+ database is fewer than three, and there are only two subjects in the SEMAINE database, we only perform the analysis of the learned representation on the DISFA database. From Fig. 6a, we find that there exist strong distributional biases in the raw feature space since the images from the same subject tend to be closer in the feature

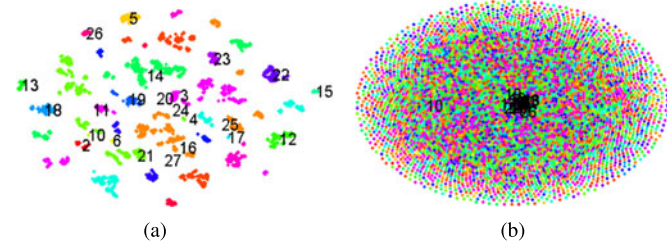


Fig. 6. (a): A t-SNE embedding of raw image feature on the DISFA database; (b): A t-SNE embedding of learned representation in terms of subjects on the DISFA database. Each text represents one subject ID and is placed at the center of its own images. The clustering effect reveals that raw face features retain individual differences; the learned representation reduces such influence.

space. However, as shown in Fig. 6b, images from the same subject tend to distribute uniformly. It demonstrates that the learned shared features diminish individual differences.

As discussed in Section 4.1, the features used in the DISFA database contain two parts: geometric feature and Gabor feature. To further analyse the effect of the learned representation on both parts, we further analyse the individual sensibility of raw features and the learned representations for geometric feature and Gabor feature. Similarly, we employ t-SNE to depict the embeddings of images represented by the geometric feature and Gabor feature, and the corresponding learned representations by the bottom hidden layer of our model. Fig. 7 lists the results. From Figs. 7a and 7b, we find that there exist individual differences among both Gabor feature and geometric feature respectively. Through the proposed model, the individual difference diminishes, as shown in Figs. 7c and 7d. Thus the

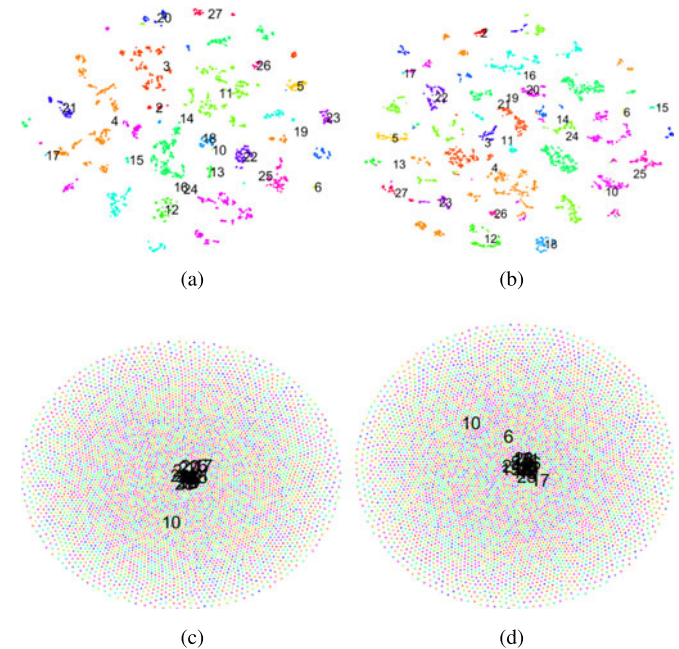


Fig. 7. The t-SNE embedding for different features in terms of subject on the DISFA database. (a): Raw gabor feature; (b): Raw geometric feature; (c): Learned representation for gabor feature; (d): Learned representation for geometric feature. Each text represents one subject ID and is placed at the center of its own images. The clustering effect reveals that raw geometric feature and gabor feature retain individual differences; the learned representation reduces such influence.

TABLE 5
AU Recognition Results Enhanced by Facial Expressions on the CK+ Database

AU	SVM	BN+[17]	HRBM+[11]	FRBM	FRBM+
1	0.938	0.945	0.925	0.923	0.924
2	0.900	0.935	0.920	0.924	0.924
4	0.779	0.785	0.784	0.779	0.794
5	0.718	0.798	0.850	0.830	0.840
6	0.756	0.776	0.798	0.822	0.798
7	0.531	0.601	0.619	0.595	0.591
9	0.875	0.911	0.900	0.915	0.944
12	0.698	0.847	0.900	0.873	0.872
17	0.811	0.837	0.850	0.854	0.828
23	0.600	0.689	0.610	0.642	0.700
24	0.268	0.447	0.600	0.582	0.625
25	0.879	0.952	0.950	0.961	0.953
27	0.852	0.977	0.895	0.901	0.908
average	0.739	0.799	0.815	0.815	0.823

learned representation has similar effect on geometric feature and Gabor feature.

4.3 Experimental Results of AU Recognition Enhanced by Facial Expression

4.3.1 Analyses of AU Recognition

Experimental results of AU recognition enhanced by facial expression on the CK+ database and the SEMAINE database are listed in Tables 5 and 6 separately. Specifically, the result of FRBM in Table 5 is obtained using the same condition with FRBM+.

On the CK+ database, we find that considering both AU relations and AU-expression relations performs better than the method only considering AU relations, since the average F1-score of FRBM+ is higher than FRBM on seven out of thirteen AUs. It suggests that the semantic relationships among AUs and expressions can facilitate AU recognition. Compared with SVM, which ignores the relations among AUs and AU-expression, our method performs better on almost all AUs. In particular, for AU24, capturing global AU dependencies among AUs improves the F1-score from 27 to 58 percent, and capturing both AU relations and AU-expression relations improves the result from 27 to 63 percent. It strongly demonstrates the effectiveness to capture the semantic relations among AUs and AU-expression. Similarly, on the SEMAINE database, the average F1-score of FRBM+ is better than FRBM and SVM, demonstrating the effectiveness of our method in dealing with spontaneous facial expressions.

We compare our work with state-of-the-art approaches which recognize AUs assisted by expressions. Wang et al. [11]

TABLE 6
AU Recognition Results Enhanced by Facial Expressions on the SEMAINE Database

AU	SVM	BN+[17]	HRBM+[11]	FRBM	FRBM+
1	0.635	0.812	0.736	0.800	0.752
2	0.687	0.760	0.717	0.752	0.713
4	0.460	0.623	0.559	0.667	0.654
5	0.364	0.400	0.400	0.400	0.606
6	0.444	0.400	0.510	0.563	0.635
7	0.353	0.039	0.373	0.506	0.468
12	0.637	0.620	0.600	0.683	0.677
17	0.125	0.182	0.462	0.364	0.292
25	0.730	0.853	0.723	0.826	0.841
26	0.592	0.607	0.237	0.676	0.684
Avg.	0.503	0.530	0.532	0.624	0.632

proposed a 3-way RBM by incorporating expression to facilitate the estimation of AU dependencies, denoted as HRBM+. Since the features and AUs used in [11] on both databases are different from ours, we re-conduct the experiments using the provided code of HRBM+. Wang et al. [17] proposed to construct a BN to capture the relationships between facial expression and AUs, denoted as BN+. They reported the results on the CK+ database, and the features and AUs used in [17] on this database are the same as ours, thus we directly compare with the reported results. For the SEMAINE database, we rerun their code using our data. From Table 5, we observe that our method performs best among three expression-augmented approaches on the CK+ database. The average F1-score of our method is 2.4 percent higher than BN+, and is 0.8 percent higher than HRBM+. Similarly, in Table 6, our method (FRBM+) performs much better than BN+ and HRBM+ on the SEMAINE database. In detail, the average F1-score of FRBM+ is about 10 percent higher than those of HRBM+ as well as BN+. This strongly demonstrates the effectiveness of our method in capturing the relations among AUs and expressions in both label space and feature space.

Due to Markov assumption, the relationships captured by BN are local. Whereas, our method captures global relations among AUs and expressions through a hidden layer. HRBM+ proposed in [11] models the global dependencies among AUs for each expression independently, while our model captures joint relations among them, which is more comprehensive.

4.3.2 Analyses of AU Dependencies Enhanced by Expressions

We take the CK+ database as an example to further analyze the benefit for AU recognition by leveraging expression-

TABLE 7
Examples of AU Recognition Results on the CK+ Database





				
expression	disgust	disgust	sadness	fear
ground-truth AU	AU4, AU7, AU9, AU25	AU4, AU6, AU7, AU9, AU25	AU1, AU4, AU17	AU1, AU4, AU25
FRBM	AU4, AU7, AU9, AU17	AU6, AU12, AU25	AU4, AU7, AU17, AU23, AU24	AU1, AU2, AU4, AU5, AU25
FRBM+	AU4, AU6, AU7, AU9, AU25	AU6, AU9, AU25	AU1, AU4, AU17	AU1, AU4, AU5, AU25

TABLE 8
The Correlation Coefficient Between Each Pair of AUs in the CK+ Database

	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU12	AU17	AU23	AU24	AU25	AU27
AU1	1.00	0.81	-0.10	0.66	-0.48	-0.39	-0.39	-0.39	-0.22	-0.22	-0.30	0.37	0.64
AU2	0.81	1.00	-0.34	0.74	-0.42	-0.37	-0.32	-0.31	-0.35	-0.20	-0.24	0.47	0.79
AU4	-0.10	-0.34	1.00	-0.22	-0.19	0.48	0.25	-0.39	0.65	0.35	0.28	-0.53	-0.39
AU5	0.66	0.74	-0.22	1.00	-0.36	-0.29	-0.28	-0.29	-0.33	-0.10	-0.24	0.47	0.65
AU6	-0.48	-0.42	-0.19	-0.36	1.00	0.16	0.00	0.69	-0.19	-0.11	-0.19	0.24	-0.34
AU7	-0.39	-0.37	0.48	-0.29	0.16	1.00	0.37	-0.22	0.41	0.33	0.33	-0.37	-0.30
AU9	-0.39	-0.32	0.25	-0.28	0.00	0.37	1.00	-0.24	0.35	-0.07	0.02	-0.39	-0.25
AU12	-0.39	-0.31	-0.39	-0.29	0.69	-0.22	-0.24	1.00	-0.39	-0.22	-0.18	0.40	-0.25
AU17	-0.22	-0.35	0.65	-0.33	-0.19	0.41	0.35	-0.39	1.00	0.38	0.32	-0.74	-0.39
AU23	-0.22	-0.20	0.35	-0.10	-0.11	0.33	-0.07	-0.22	0.38	1.00	0.52	-0.42	-0.18
AU24	-0.30	-0.24	0.28	-0.24	-0.19	0.33	0.02	-0.18	0.32	0.52	1.00	-0.43	-0.21
AU25	0.37	0.47	-0.53	0.47	0.24	-0.37	-0.39	0.40	-0.74	-0.42	-0.43	1.00	0.48
AU27	0.64	0.79	-0.39	0.65	-0.34	-0.30	-0.25	-0.25	-0.39	-0.18	-0.21	0.48	1.00

TABLE 9
The Number of Samples That Each AU Appears in Each Expression

	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU12	AU17	AU23	AU24	AU25	AU27
anger(45)	0	0	40	6	8	32	3	1	39	36	33	0	0
contempt(18)	1	1	1	0	0	0	0	5	5	1	2	0	0
disgust(59)	0	0	36	0	18	33	58	2	41	2	7	9	0
fear(25)	22	10	21	16	3	6	0	2	3	0	0	23	0
happy(69)	0	0	0	0	66	7	0	67	0	0	0	67	0
sadness(28)	26	7	23	0	0	1	0	0	27	3	1	0	0
surprise(83)	81	81	1	70	0	0	0	3	0	1	0	82	72

The digit in bracket represents the total number of samples of each expression.

dependent AU relations. Table 8 lists the correlation coefficients between each pair of AUs on the CK+ database. Positive value represents a positive correlation, and negative value represents a negative correlation. It depicts the global co-occurrence and mutual-exclusive relations among multiple AUs. Table 9 lists the occurrence frequency of AUs for each expression. It depicts the relations among AUs and expressions. To further demonstrate the effectiveness of exploiting the expression-related relations among AUs, we list several examples in Table 7, which are obtained from the CK+ database. For example, for a sample labeled with AU4, AU7, AU9 and AU25, both FRBM and FRBM+ correctly predict AU4, AU7 and AU9. These three AUs are highly positively correlated with each other, since the correlation coefficients among these AUs in Table 8 are higher than zero. And our proposed two methods are able to capture these co-occurrence relations. Moreover, FRBM+, which uses expression to assist AU recognition, further infers the occurrence of AU25, which is negatively correlated with AU4, AU7 and AU9, as shown in Table 8. However, from Table 9, we find that for disgust expression there are some cases that AU25 appears along with AU4, AU7, AU9. It demonstrates the expression-related relation among AUs is more stronger. Similarly, for a disgust face appears with one more AUs (AU6), FRBM+ can recognize AU9 which is negatively correlated with AU6 and AU25, while FRBM predicts AU6, AU12 and AU25, which are more likely to occur together. For another instance, for a sample with the presence of AU1, AU4 and AU17, the results of FRBM are AU4, AU7, AU17, AU23 and AU24, but without AU1, in that there exists a positive correlation among each

pair of the former five AUs, while a negative correlation between AU1 and AU4/AU17. In contrast, the results of FRBM+ enhanced by expression are on target, since in most cases AU1 and AU4 appear together on a sad face, as shown in Table 9. From Table 8, it can be observed that AU1 and AU2 are more likely to occur together. So when recognizing an image with the presence of AU1, like the last column in Table 7, FRBM is much likely to target with both AU1 and AU2. Nevertheless, an expression assisted method of FRBM+ performs well. These confirm that the co-occurrence and mutual exclusive relations among AUs are related to expressions. As discussed in Section 3, the top hidden layer in Figs. 2a and 2b captures a different pattern. In Fig. 2a, the hidden layer captures a high-order dependencies among AUs. While in Fig. 2b, it can also model the global relations between AUs and expressions, which is much stronger.

4.3.3 Analyses of Expression Recognition

Technically our proposed model FRBM+ can perform AU recognition and expression classification simultaneously. To analyse the effect of our model in jointly modeling, we further conduct expression recognition experiment. The results are listed in Table 10. From this table, we find that

TABLE 10
The Results of Expression Recognition Accuracy

Database	FRBM+	related work
CK+	0.880	0.863 [25]
SEMAINE	0.580	-

the expression classification accuracy on the CK+ database is 0.88, which is comparable with that of the state-of-the-art approach [25]. It demonstrates the superiority of our approach. The result on the SEMAINE database is 0.58. To the best of our knowledge, there is no work reported the expression recognition results on the SEMAINE database.

5 CONCLUSION AND FUTURE WORK

In this paper, a multiple AU recognition approach is proposed by learning relations from both feature level and label level. We employ a hierarchical RBM model, which includes two hidden layers. The top hidden layer is imposed on AUs to model global semantic relationships among multiple AUs. The lowest hidden layer embedded between AUs and features aims at capturing the commonalities and characteristics among different AUs. The hierarchy RBM model is then augmented with the facial expression labels during training to further improve AU recognition performance. Efficient learning and inference algorithms are introduced for both models. The experimental results on several benchmark databases demonstrate that both FRBM and FRBM+ can successfully learn joint feature and label relations for improving multiple AU recognition.

The proposed FRBM+ requires both expression and AU labels during training. To make FRBM+ applicable with only AU labels or expression labels during training, the prior knowledge on statistical correlations between AUs and expressions can be used to infer expression labels from AU labels or generate AU labels from expression labels. This is our future work.

ACKNOWLEDGMENTS

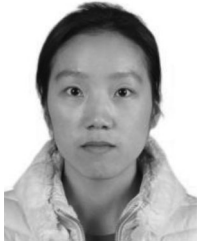
This work has been supported by the National Science Foundation of China (grant nos. 61473270, 61175037 and 61228304) and the project from Anhui Science and Technology Agency (1508085SMF223).

REFERENCES

- [1] P. Ekman, "Facial expression and emotion," *Amer. Psychologist*, vol. 48, no. 4, 1993, Art. no. 384.
- [2] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
- [3] Z. Zeng, M. Pantic, G. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
- [4] S. Wang, et al., "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.
- [5] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3D/4D facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1438–1450, Jul. 2016.
- [6] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1319–1329, Jul. 2016.
- [7] E. Friesen and P. Ekman, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Psychologists Press, 1978.
- [8] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.
- [9] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, Apr. 2016.
- [10] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.
- [11] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3304–3311.
- [12] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor, "Exploiting sparsity and co-occurrence structure for action unit recognition," in *Proc. 11th IEEE Int. Conf. Workshops Automatic Face Gesture Recognit.*, 2015, pp. 1–8.
- [13] Y. Zhu, S. Wang, L. Yue, and Q. Ji, "Multiple-facial action unit recognition by shared feature learning and semantic relation modeling," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1663–1668.
- [14] X. Zhang and M. H. Mahoor, "Simultaneous detection of multiple facial action units via hierarchical task structure learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1863–1868.
- [15] K. Zhao, W.-S. Chu, F. De la Torre Frade, J. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2207–2216.
- [16] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3792–3800.
- [17] J. Wang, S. Wang, and Q. Ji, "Facial action unit classification with hidden knowledge under incomplete annotation," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 75–82.
- [18] A. Ruiz, J. Van de Weijer, and X. Binefa, "From emotions to action units with hidden and semi-hidden-task learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3703–3711.
- [19] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 94–101.
- [21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr.-Jun. 2013.
- [22] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan.-Mar. 2012.
- [23] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn, "A 1 p-norm MTMML framework for simultaneous detection of multiple facial action units," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 1104–1111.
- [24] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *Automat. Face Gesture Conf.*, vol. 4, 2017.
- [25] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3422–3429.



Shangfei Wang received the BS degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996, the MS degree in circuits and systems, and the PhD degree 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 with Kyushu University, Japan. Between 2011 and 2012, she was a visiting scholar at Rensselaer Polytechnic Institute in Troy, New York. She is currently an associate professor of School of Computer Science and Technology, USTC. Her research interests cover affective computing, and probabilistic graphical models. She has authored or co-authored more than 90 publications. She is a senior member of the IEEE and a member of the ACM.



Shan Wu received the BS degree in computer science from the Anhui University of Technology, in 2014, and she is currently working toward the MS degree in computer science at the University of Science and Technology of China, Hefei, China. Her research interest is affective computing.



Guozhu Peng received the BS degree in mathematics from the South China University of Technology, in 2016, and he is currently working toward the MS degree in computer science at the University of Science and Technology of China, Hefei, China. His research interest is affective computing.



Qiang Ji received the PhD degree in electrical engineering from the University of Washington. He is currently a professor in the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI). He recently served as a program director at the US National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute, University of Illinois at Urbana-Champaign, the

Robotics Institute at Carnegie Mellon University, the Department of Computer Science, University of Nevada at Reno, and the US Air Force Research Laboratory. He currently serves as the director of the Intelligent Systems Laboratory (ISL), RPI. His research interests include computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published more than 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including US National Science Foundation, NIH, DARPA, ONR, ARO, and AFOSR as well as by major companies including Honda and Boeing. He is an editor on several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. He is a fellow of the IAPR and the IEEE.

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