Deep Facial Action Unit Recognition from Partially Labeled Data

Shan Wu[1], Shangfei Wang*[,1], Bowen Pan[1], and Qiang Ji[2]
[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
SA14WS@mail.ustc.edu.cn, sfwang@ustc.edu.cn, bowenpan@mail.ustc.edu.cn, qji@ecse.rpi.edu

Abstract

Current work on facial action unit (AU) recognition requires AU-labeled facial images. Although large amounts of facial images are readily available, AU annotation is expensive and time consuming. To address this, we propose a deep facial action unit recognition approach learning from partially AU-labeled data. The proposed approach makes full use of both partly available ground-truth AU labels and the readily available large scale facial images without annotation. Specifically, we propose to learn label distribution from the ground-truth AU labels, and then train the AU classifiers from the large-scale facial images by maximizing the log likelihood of the mapping functions of AUs with regard to the learnt label distribution for all training data and minimizing the error between predicted AUs and ground-truth AUs for labeled data simultaneously. A restricted Boltzmann machine is adopted to model AU label distribution, a deep neural network is used to learn facial representation from facial images, and the support vector machine is employed as the classifier. Experiments on two benchmark databases demonstrate the effectiveness of the proposed approach.

1. Introduction

Facial behavior is one of the most important emotion communication channels for human-human interaction. Therefore, a great progress on automatical facial action unit recognition has been achieved due to its wide application in many user-centered fields in recent years.

Automatical AU recognition is very challenging due to many factors, such as imaging conditions and individual subject differences. Large-scale training data can facilitate the learning process of AU classifiers. Although it is easy to obtain a large-scale facial images due to the popularity of digital cameras, portable devices and internet, providing AU labels for the obtained large-scale facial images is very time consuming and difficult, since AU labels should be annotated by experts. However, almost all current work formulates AU recognition as a supervised learning process, and thus requires facial images and their corresponding AU labels. The requirement of corresponding AU labels prevents leveraging large-scale available facial images for improving AU recognition.

Fortunately, behavior research shows that there exist regular spatial and temporal patterns in AU labels due to facial anatomy and human's behavior habits. For example, as stated in Du *et al.*'s [2] work, 99 percent of the time, persons show happiness by raising their cheeks and stretching their mouth in a smile. Persons can not pull lip corner (AU12) and depress lip corner (AU15) at the same time due to the constraint of facial anatomy. Such regular spatial and temporal patterns embedded in AU labels can be described probabilistically as label distribution, which is also existed in ground-truth AU labels of existing benchmark expression databases. Take the DISFA database as an example, as shown in Figure 1, when AU6 appears, AU12 and AU25 occur with a probability higher than 0.7, but AU1 rarely occurs. Similarly, when AU12 or AU25 occur, AU6 appears with a probability higher than 0.6. Therefore, the annotated AU labels are samples of AU label distributions, which are inherent in facial anatomy and human's behavior habits. In fact, the AUs appear on any facial images, whether manually annotated or not, are samples of the underlying AU label distributions. Inspired by the above observations, we propose a deep facial action unit recognition approach learning from partially AU-labeled training data through incorporating such spatial regular patterns of AU labels presented in

	1	2	4	6	9	12	15	17	25	26
1	1.00	0.64	0.45	0.18	0.16	0.20	0.13	0.19	0.55	0.35
2	0.90	1.00	0.35	0.12	0.04	0.16	0.07	0.10	0.41	0.26
4	0.30	0.16	1.00	0.37	0.43	0.15	0.31	0.39	0.55	0.31
6	0.09	0.04	0.27	1.00	0.22	0.74	0.11	0.13	0.90	0.40
9	0.21	0.04	0.85	0.60	1.00	0.21	0.28	0.38	0.71	0.36
12	0.09	0.05	0.11	0.71	0.07	1.00	0.02	0.06	0.94	0.46
15	0.21	0.08	0.75	0.36	0.35	0.08	1.00	0.56	0.49	0.30
17	0.26	0.10	0.80	0.37	0.40	0.18	0.48	1.00	0.29	0.21
25	0.17	0.09	0.27	0.61	0.18	0.66	0.10	0.07	1.00	0.64
26	0.20	0.11	0.27	0.48	0.16	0.58	0.11	0.09	0.97	1.00

Figure 1. Dependencies between AU labels (each entry represents the conditional probability of $p(y_j = 1|y_i = 1)$)

*This is the corresponding author.

ground-truth AU labels into the learning process of AU classifiers from a large-scale facial images without AU annotations. Specifically, we utilize the Restricted Boltzmann Machine (RBM) [5] to learn the AU label distribution from the available AU labels. And a large number of facial images are used to learn a deep framework to exploit the facial representation from facial images. Then we train multiple support vector machines through maximizing the log likelihood of the mapping functions of AUs with regard to the learnt label distribution for all training data and minimizing the error between predicted AUs and ground-truth AUs for labeled data simultaneously. The experimental results on two benchmark databases prove the effectiveness of deep neural network in feature learning and illustrate that the AU label constraint can improve the performance of AU recognition under either complete AU annotations or incomplete AU annotations.

2. Related work

2.1. AU recognition using shallow models

Most AU recognition works classify AUs from the handcraft features, i.e., geometric features and appearance features. We refer to them as shallow models. In addition to recognizing each AU independently, recent work begins to exploit AU dependencies for multiple AU recognition through either generative strategy or discriminative strategy.

For generative approaches, the structure and parameters of probabilistic graphic models are adopted to capture AU dependencies from AU labels. For instance, Tong *et al.* [16] introduced a generative model based on dynamic Bayesian networks (DBN) to model the semantics of different AUs through its structure and conditional probabilities. Wang *et al.* [18] proposed a three-layer RBM model to capture high-order dependencies among different AUs, and the weights between hidden layer and target labels measure the probabilistic patterns among AU labels. These two works learned AU relations from ground-truth AU labels. However, Li *et al.*[21] proposed to generate pseudo-data according to AU dependencies summarized from prior knowledge, then learnt a Bayesian Network (BN) model from the pseudo-data.

For discriminative approaches, the relations among AUs are adopted as constraint of the objective function. For example, Zhu *et al.* [26] and Zhang *et al.* [22] used the constraint to supervise the multi-task learning process so as to exploit the AU co-occurrences among AU labels and facial regions. Zhao et al. [24] leveraged group sparsity by selecting a sparse subset of facial patches while learning a multi-label classifier. Through jointly patch learning, both positive correlations and negative competitions among AUs are introduced to model a discriminative multi-label classifier. Eleftheriadis et al. [4] proposed a multi-conditional latent variable model by performing the fusion of different facial features and AU detection jointly. They attained the feature fusion by learning a low-dimensional subspace which is constrained by the local dependencies among multiple AUs.

All of the above work requires complete AU annotations to train AU classifiers. It is not practical since AU annotation is an expensive and time consuming task. What's more, all of them use hand-craft features, and thus do not fully exploit the current development of deep learning and large-scale facial images.

2.2. AU recognition from partially annotated training sample

The mainstream of AU recognition trains AU classifiers from fully annotated facial images. Only very recently, a few works begins to consider AU classification under partial AU annotations.

Wang *et al.* [17] proposed an expression-assisted AU recognition method under incomplete AU labeling. A BN model is adopted to capture the dependencies among AUs and expression, and a structured EM is used to learn the structure and parameters of the BN when AU labels are missing. Although Wang *et al.*' work can handle incomplete AU labeling, their method requires expression labels as hidden knowledge to complement the missing AU labels.

Wu *et al.* [20] formulated AU recognition under partial AU annotations as a multi-label learning with missing labels (MLML). They proposed to handle the missing labels by enforcing the consistency between the predicted labels and the provided labels as well as the local smoothness among the label assignments. The same features are used for all AU classifiers. Since the discriminative features for each AU are different, Li *et al.* [11] extended Wu *et al.*'s MLML method to discriminate each AU based on the most related features. Both work assumes local smoothness to handle missing labels. However, this assumption may be invalid, since the samples closed in feature space may belong to the same subject, other than the same expressions.

Song *et al.* [15] proposed a Bayesian graphical model that simultaneously handles sparsity and co-occurrence structure of facial action units using compressed sensing and group wise inducing priors. Their method handles missing labels by marginalizing over the unobserved values during the inference procedure. As a generative model, Song *et al.* [15] work can hand missing labels effectively, but may not obtain better performance in classification tasks compared with discriminative models.

Ruiz *et al.* [13] proposed to learn AU classifiers under the help of a large-scale facial images with expression labels, but without AU labels. They proposed to exploit prior knowledge about the relations between Hidden-Tasks (AUs) and Visible-Tasks (expressions). Although this works can

learn AU classifiers from images without AU labels, it requires another big amount of expression-annotated facial images.

Although these work considers AU recognition under incomplete AU annotations, all of them used hand-craft features. In order to handle missing annotations, they adopted local smoothness, or required expression labels, or employed generative models. While, in our paper, we propose a discriminative model to exploit the inherent label distributions, which are extracted from ground-truth AU labels and are caused by facial anatomy and human behavior habits, as weak supervisory information to help feature learning and AU recognition with the deep model.

2.3. AU recognition using deep networks

Recently, feature learning using deep hierarchical structures has emerged as an effective methodology to automatically extract features from data for many computer vision problems. Most deep networks for AU detection directly adopt Convolutional Neural Network (CNN) to learn spatial representation. Ghosh et al. [6] proposed a multi-label convolutional neural network approach to learn a shared representation between multiple AUs from facial images. Gadi et al. [7] proposed to estimate AU occurrence and intensity using a 7-layer CNN, which consists of 3 convolutional layers and a max-pooling layer. Khorrami et al.[10] examined how much convolutional neural networks (CNNs) can improve performance on expression recognition and what they actually learn. Their experimental results showed that CNNs trained for expression recognition are indeed able to model high-level features that strongly correspond to AUs. Zhao et al. [25] proposed Deep Region and Multi-label Learning (DRML), a unified deep network that simultaneously addresses region learning and multi-label learning for multiple AU recognition. The proposed DRML architecture consists of a standard convolution layer filtering on an aligned face image, followed by the region layer, one pooling layer and four convolution layers, three fully connected layers, and one multi-label cross-entropy loss layer at the end. To address the overfitting problem of CNN due to limited training data, Han et al. [14] proposed an incremental Boosting CNN (IB-CNN) to integrate boosting into the CNN through introducing an incremental boosting layer and a new loss function. The incremental boosting layer selects discriminative neurons from the lower layer and is incrementally updated on successive mini-batches. The loss function consists of errors from both the incremental boosted classifier and individual weak classifiers. Other than using CNN to capture spatial representation only, Jaiswal and Valstar [9] combined convolutional neural networks and bi-directional long short-term memory neural networks (CNN-BLSTM) to jointly learn shape, appearance and dynamics in a deep learning manner for AU recognition.

All these work demonstrates the potential of deep network for AU recognition. However, all of them require a large number of fully annotated facial images, which is unrealistic and is very time consuming.

To sum up, current research on AU recognition either uses hand-craft features or recognizes AUs under fully labeled data. Therefore, in this paper, we propose a method to perform AU recognition with partially labeled data based on a deep framework. Specifically, we learn a deep neural network from a large amount of images, and then learn the AU label distribution from partially available AUs. During the learning process of AU classifiers, we optimize the model parameters with the help of the learnt AU label distribution.

Compared with related work, our contributions are as follows: (1) we are the first to deal with AU recognition with partially labeled data using a deep framework by exploiting the AU label distribution during the training phase of AU classifier; (2) we perform experiments with completely labeled data and partially labeled data, demonstrating the effectiveness of AU label distribution constraint.

3. Problem Statement

The purpose of our work is to learn AU classifier from a larger-scale partially-labeled facial images. Our assumption is that the regular spatial patterns embedded in AU labels can be described probabilistically as label distribution. Both the ground-truth AU labels for labeled facial images and the un-annotated AU labels for the larger-scale unlabeled facial images are samples of such distributions. Therefore, in addition to leaning AU classifiers by minimizing label errors between predicted AU labels and the ground-truth AU labels, we regulate the AU classifier through maximizing the log likelihood of the mapping functions of AUs with regards to the label distribution, which can be learnt from the ground-truth AU labels.

Let $D = \{\mathbf{x}_t, \mathbf{y}_t\}_{t=1}^{N}$ denotes a set of d dimensional training instances $\mathbf{x}_t \in \mathbb{R}^d$ and the corresponding labels $\mathbf{y}_t \in \{-1, 0, 1\}^p$, where p is the number of labels and N is the number of instances. Specifically, the positive label is denoted as 1, the negative label is denoted as -1, and the missing label is denoted as 0. Therefore, $\mathbf{y}_t \in \{-1, 1\}^p$ represents the case of completely labeled examples, and $\mathbf{y}_t \in \{-1, 0, 1\}^p$ represents the case of partially labeled examples. Given the training data D, our goal is to learn a classifier $f : \mathbb{R}^d \rightarrow \{-1, 1\}^p$ from partially labeled data according to Equation 1. The first term in this equation is the error between predicted AUs and ground-truth AUs for labeled data. The second term is the log likelihood of the mapping functions of AUs with regard to the label distribution learnt from AU label set, and this term acts as a constraint to make the output of $f(\mathbf{x})$ to be more approximate

with the regular spatial patterns embedded in AU labels.

$$\min_{\Theta} \sum_{t=1}^{N} Loss(\mathbf{y}_t, f(\mathbf{x}_t; \Theta)) - \beta \sum_{t=1}^{N} \log p(f(\mathbf{x}_t; \Theta)) \quad (1)$$

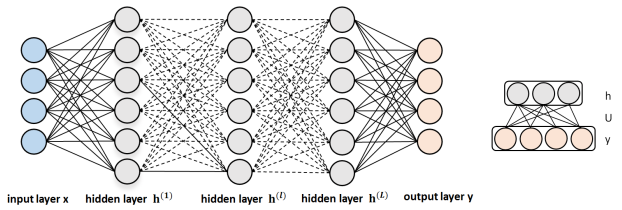
As can be seen in Equation 1, when $\beta = 0$, the optimization problem is equivalent to learning the AU classifiers without taking the AU label distribution into account; otherwise, our model can learn the classifier either with completely annotated images in a supervise way or partially labeled data in a semi-supervise manner.

4. Proposed approach

As discussed in Section 1, AU annotation is a time consuming task which needs professionals to annotate multiple AUs for each image. Therefore there is a serious lack of images annotated with AU labels for AU recognition research. However, there are a large amount of images without AU labels, thus we can utilize a deep learning method to train a deep framework to learn some important properties of each image. Furthermore, there exists mutual-exclusive and co-occurrence relations among multiple AUs due to the facial muscle structure. Restricted Boltzmann Machine is a good model to capture the inherent patterns in visible units. Thus we use this model to capture the inherent relations among multiple AU labels to improve the AU recognition results. In this section, we introduce the details of our proposed model.

4.1. Deep Neural Network

Deep neural network (DNN) is a feedforward network with many hidden layers. As shown in Figure 2(a), $\mathbf{x}$ represents input features (usually the raw features, such as the raw pixel of image), $\mathbf{h}^{(l)}$ represents the l-th hidden layer, and $\mathbf{y}$ represents the target outputs, e.g. AUs. In the DNN model, each two adjacent layers can be considered as a sharing component. Thus we can train the deep network in a layer-wise manner. Through multiple hidden layers, the raw input features $\mathbf{x}$ can be encoded with a high-order representation $\mathbf{h}^{(L)}$. The top hidden layer $\mathbf{h}^{(L)}$ and the output layer $\mathbf{y}$ form a classifier. In our work, we employ multiple Support Vector Machines (SVM) for multiple AU recognition based on the high-layer features, i.e., $\mathbf{h}^{(L)}$. Thus AU recognition is performed according to Equation 2, where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_p), \mathbf{d} = (d_1, \ldots, d_p)$ is the parameters for p classifiers of SVM, and $\mathbf{w}_i \in \mathbb{R}^d, d_i \in \mathbb{R}$ is the parameters for the i-th SVM.



(a) DNN (b) RBM

Figure 2. The structure of DNN and the target prior model (RBM).

$$y_i = sign(\mathbf{w}_i^T \mathbf{h}^{(L)} + d_i) \quad (2)$$

The loss function is defined in Equation 3, where the first item is a L2 regularization term, the second item is the squared hinge loss function, and α is a hyper-parameter. Different with the original hinge loss function, we replace the constant 1 by $\mathbf{y}^T\mathbf{y}$ in order to make the loss function also apply to the case of missing tags. When y_i is not missing, i.e., $y_i = 1$ or $y_i = -1$, $y_i \times y_i = 1$; when y_i is missing, i.e., $y_i = 0$, the hinge loss is equal to 0, which does not increase the value of the loss function.

$$L(\Theta) = \frac{1}{2}\|\mathbf{W}\|_2^2 + \alpha \sum_{t=1}^{N} \left\| \left[\mathbf{y}_t \cdot \mathbf{y}_t - \mathbf{y}_t \cdot \left(\mathbf{W}^T \mathbf{h}_t^{(L)} + \mathbf{d} \right) \right]_+ \right\|_2^2 \quad (3)$$

4.2. AU Label Distribution

A Restricted Boltzmann Machine (RBM) is bipartite network, including two types of nodes: visible units and hidden units. A graphical depiction of RBM is shown in Figure 2(b), where $\mathbf{h}$ represents the hidden units, and $\mathbf{y}$ represents the visible units. In our work, we use RBM to model the global dependencies among multiple AUs.

The energy function of RBM is defined as follows,

$$E(\mathbf{h}, \mathbf{y}; \Upsilon) = -\mathbf{b}^T\mathbf{y} - \mathbf{c}^T\mathbf{h} - \mathbf{y}^T\mathbf{U}\mathbf{h} \quad (4)$$

where $\Upsilon = (\mathbf{U}, \mathbf{b}, \mathbf{c})$ represents the parameters of the model. Specifically, b_i represents the bias of AU labels, c_i represents the bias of hidden units, and U_{ij} represents the weight between the i-th AU and j-th hidden unit.

The joint distribution $p(\mathbf{y})$ of the model is defined as follows,

$$p(\mathbf{y}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{h}, \mathbf{y}; \Upsilon)) \quad (5)$$

where $Z = \sum_{\mathbf{y},\mathbf{h}} \exp(-E(\mathbf{h}, \mathbf{y}; \Upsilon))$.

When analysing Equation 4 and Equation 5, we can find that there is a positive correlation between $p(\mathbf{y})$ and $\mathbf{U}$. In more detail, the larger weight represents a higher probability of AU occurrence, and the smaller weight represents a higher probability of AU absence. In other words, the weights between the hidden units and the AUs capture the mutual-exclusive and co-occurrence relationships among AU labels.

According to Equation 6, we employ a Maximum Likelihood Estimation method to learn the parameters Υ. Due to the complexity of calculating the normalizing factor Z, we use contrastive divergence [8] to estimate Υ. The gradient of $\gamma \in \Upsilon$ is defined in Equation 7.

$$\Upsilon^* = \underset{\Upsilon}{argmax} \log p(\mathbf{y}; \Upsilon) \quad (6)$$

$$\frac{\partial \log p(\mathbf{y})}{\partial \gamma} = \langle \frac{\partial E}{\partial \gamma} \rangle_{p(\mathbf{h},\mathbf{y})} - \langle \frac{\partial E}{\partial \gamma} \rangle_{p(\mathbf{h}|\mathbf{y})} \tag{7}$$

4.3. Learning with AU label constraint

Since the DNN learning process ensures fidelity in data reconstruction, such learnt features may not perform well on a specific task. What's more, there exists co-occurrence and mutual-exclusive relations among AU labels. Hence we intend to capture the inherent relations among AU labels to regularize the feature learning by making the prediction of the target variable consistent with the AU label distribution [19]. Due to the range of visible units in RBM model, different with the previous setting, we set the negative label to 0.

The AU label constraint is defined in Equation 8. When parameters $\Upsilon = (\mathbf{U}, \mathbf{b}, \mathbf{c})$ is fixed, the normalizing factor $\mathbf{Z}$ in $p(\mathbf{y})$ is a constant, so we drop it off from the constraint. The derivation of $\log \tilde{p}(\mathbf{y})$ over $\mathbf{y}$ is given by Equation 9, where $\sigma(s) = \frac{1}{1+exp(-s)}$ is a sigmoid function. Thus the derivation of $\log \tilde{p}(\mathbf{y})$ over Θ is obtained by $\frac{\partial \log \tilde{p}(\mathbf{y})}{\partial \Theta} = \frac{\partial \log \tilde{p}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \Theta}$.

$$\begin{aligned} \log \tilde{p}(\mathbf{y}) &= \log \sum_{\mathbf{h}} \exp(-E(\mathbf{y}, \mathbf{h})) \\ &= \sum_i y_i b_i + \sum_j \log(1 + \exp(c_j + \sum_i y_i U_{ij})) \end{aligned} \tag{8}$$

$$\frac{\partial \log \tilde{p}(\mathbf{y})}{\partial \mathbf{y}} = \mathbf{b} + \sum_j \sigma(c_j + \sum_k y_k U_{kj}) \mathbf{U}_{\cdot j} \tag{9}$$

Considering the AU label distribution constraint, given the training data $D = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^N$, the loss function is defined as follows,

$$\begin{aligned} F(\Theta) =& \frac{1}{2}\|\mathbf{W}\|_2^2 + \alpha \sum_{t=1}^N \left\| \left[\mathbf{y}_t \cdot \mathbf{y}_t - \mathbf{y}_t \cdot \left(\mathbf{W}^T \mathbf{h}_t^{(L)} + \mathbf{d} \right) \right]_+ \right\|_2^2 \\ & - \beta \sum_{t=1}^N \log \tilde{p} \left(\sigma \left(\mathbf{W}^T \mathbf{h}_t^{(L)} + \mathbf{d} \right); \Upsilon \right) \end{aligned} \tag{10}$$

where $\alpha \geq 0, \beta \geq 0$ are the hyperparameters , and Θ includes the parameters of both DNN and SVM.Υ is already learnt from the available AU labels. Minimizing $F(\Theta)$ is equivalent to minimizing the structural risk for SVM and maximizing the log likelihood of the AU label distribution model.

4.3.1 Parameters Learning

As discussed in the previous section, the AU label distribution can be learnt from the available AU labels. Given a set of training data $D = \{\mathbf{x}_t, \mathbf{y}_t\}_{t=1}^N = \{\mathbf{X}, \mathbf{Y}\}$, we learn our model from the image data $\mathbf{X}$ and the associated AU labels $\mathbf{Y}$. First, we pre-train the DNN based on the image data in an unsupervised manner, and then fine-tune the parameters of both DNN and SVM by minimizing the objective function $F(\Theta)$ defined in Equation 10.

As explained in Section 4.1, we can use partially labeled data during the procedure of fine-tuning. From Equation 10, we find that the completely labeled samples supervise the model learning in two ways: minimizing the differences between the ground-truth labels and the predicted labels, and making the predicted labels to be consistent with the learnt AU label distributions. As for the samples without AU labels, they have no contribute on the hinge loss function, but play a role in AU label constraint. Therefore, our proposed model can not only utilize the unlabeled data to pre-train the DNN, but also can use them to fine-tune the parameters of DNN and AU classifiers through the AU label constraint.

During the fine-tuning process, we employ a backpropagation method to learn the parameters, which uses a chain rule to calculate the gradient of each parameter. We define the derivation of $F(\Theta)$ over the weighted sum of inputs in the l-th hidden layer $\mathbf{z}^{(l)}$ as $\boldsymbol{\delta}^{(l)}$, which can be calculated according to Equation 11 in a backward propagation way.

$$\boldsymbol{\delta}^{(l)} = \frac{\partial F(\Theta)}{\partial \mathbf{z}^{(l)}} = \left((\mathbf{W}^{(l+1)})^T \boldsymbol{\delta}^{(l+1)} \right) \cdot \mathbf{h}^{(l)} \cdot \left(1 - \mathbf{h}^{(l)} \right) \tag{11}$$

Thus the gradient of the weight $\mathbf{W}^{(l)}$ of the DNN model can be obtained according to Equation 12 (to simplify, we only consider the gradient of one sample). Specifically, $\mathbf{h}^{(0)}$ is the image features $\mathbf{x}$.

$$\frac{\partial F(\Theta)}{\partial \mathbf{W}^{(l)}} = \boldsymbol{\delta}^{(l)} \left(\mathbf{h}^{(l-1)} \right)^T, l = 1, \ldots, L \tag{12}$$

Assuming $\mathbf{s} = \mathbf{W}\mathbf{h}^{(L)} + \mathbf{d}$, we obtain the derivation of $F(\Theta)$ over the parameter $\mathbf{W}$ of SVM according to Equation 13.

$$\begin{aligned} \frac{\partial F(\Theta)}{\partial W_{ij}} =& W_{ij} - 2\alpha h_i^{(L)} y_j \left[y_j^2 - y_j \left(\sum_i h_i^{(L)} W_{ij} + d_j \right) \right]_+ \\ & - \beta \frac{\partial \log \tilde{p}(\sigma(\mathbf{s}))}{\partial \sigma(s_j)} \sigma(s_j) \left(1 - \sigma(s_j)\right) h_i^{(L)} \end{aligned} \tag{13}$$

4.3.2 Inference

Given a query sample $\hat{\mathbf{x}}$, first, we obtain the state of top hidden layer $\hat{\mathbf{h}}^{(L)}$ by using a forward propagation, then we perform AU recognition through the learnt multi-SVMs by Equation 2.

5. Experiments

5.1. Experimental Conditions

Two spontaneous databases are used in our experiments: the BP4D database [23] and the DISFA database [12].

The BP4D database contains 2D/3D videos of spontaneous facial expressions in young adults during various emotion inductions while interacting with an experimenter. It provides 328 2D videos with 12 AUs coded from 41 participants, resulting in $\sim 140,000$ valid samples. In our work, we use all the valid samples and all the AUs.

The DISFA database contains video recordings of 27 subjects while watching YouTube videos. Each image frame was coded in terms of 12 AUs and each AU intensity is ranked from 0 to 5. In our work, we select samples whose summation of all AUs' intensity is larger than 6, resulting in 19850 samples. Then, 10 AUs whose frequency of occurence is higher than 10% are employed in our work, i.e., AU1, AU2, AU4, AU6, AU9, AU12, AU15, AU17, AU25, AU26. And we treat each AU with intensity larger than zero as active.

The images in both databases are normalized to 100×100. On the BP4D database, we randomly divide the data into three parts according to subjects: 60% subjects for training, 20% subjects for validating, and the rest for testing. In order to reduce the impact of randomness, each experiment on this database is performed for five times, and the mean F1 Score is used to evaluate the experimental results. For the DISFA database, a 10-fold subject-independent cross validation is adopted due to the relatively small number of samples.

To demonstrate the effectiveness of our method, two experiments were conducted: AU recognition under complete AU annotation and AU recognition under incomplete AU annotation. For complete AU annotation, we compare our proposed method constrained by AU label distribution with the method without AU label constraint. Similarly, for incomplete AU annotation, we compare our method with AU label constraint with the method without AU label distribution constraint. For the experiment under incomplete AU annotation, we randomly miss the AU labels with certain proportion, i.e., 10%, 20%, 30%, 40%, 50%.

5.2. Experimental Results and Analysis

5.2.1 AU recognition with completely annotated data

1) Experimental results: Table 1 and Table 2 summarize the experimental results of AU recognition with AU label distribution constraint under completely annotated data on the BP4D database and the DISFA database respectively. As can be seen from these two tables, employing AU label distribution to regularize the AU recognition can improve the AU detection performance. On the BP4D database, the results of our method constrained by AU label distribution are higher than those without AU label constraint on 9 out of 12 AUs, and the average F1 Score of ours is about 5% higher as well. It strongly demonstrates the effectiveness of AU label distribution on regularizing the AU recognition. Similarly, on the DISFA database, our method performs better in most cases, and the average F1 Score of ours is 5% higher than the method without AU label constraint. In particular, there is a significant improvement on AU 17, which is improved from 0.201 to 0.447. What's more, AU15 and AU6 are improved for about 10%. These further illustrate the effectiveness of probabilistic relations among different AUs. *2) Evaluation of learnt representation:* To validate the effectiveness of DNN, we employ t-SNE to depict the embedding of images represented by the raw pixel of images and the learnt deep representation, and visualize the effect of individual differences by coloring in terms of subjects [1]. Figure 3 lists the examples of AU12 on the BP4D database and AU25 on the DISFA database. From Figure 3(a) and Figure 3(c), we can see that there exist strong distributional biases in the raw face images since the images from the same subject tend to be closer in the feature space. However, as shown in Figure 3(b) and Figure 3(d), images from the same subject tend to distribute uniformly. It demonstrates that the deep neural network diminishes the individual differences. *3) Comparison with related work:* To further evaluate the superiority of our method under complete AU annotation, we compare our proposed approach with state-of-the-art learning approaches for AU recognition. [25] compared their work with many state-of-the-art methods on the BP4D database and the DISFA database, including deep models and shallow models. However, the results on the DISFA in [25] were reported using the model learnt on the BP4D database, therefore we only compare our work with theirs on the BP4D database only. [3] and [4] provided experimental results on the DISFA database, thus we directly compare our results on the DISFA database with those in [4] and [3].

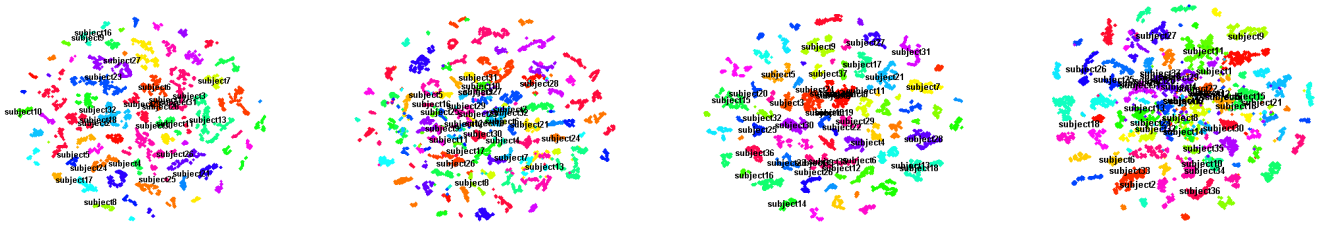
From Table 1, we can find that the average F1 Scores of our method are higher than the results of state-of-the-art approaches. Compared with the deep learning methods, i.e., DRML, AlexNet, ConvNet and LCN, our method performs best. In more detail, the average F1 Score of ours is 11% higher than AlexNet, 2% higher than ConvNet, and 3% higher than LCN, respectively. The above three work did not consider the AU relationships, while our method regularizes the classifier training and feature learning process by AU label distribution constraint. This leads to a better performance. Compared with DRML, the results of our method are better on 7 AUs. DRML introduces a region layer that uses feed-forward functions to capture structural information in different facial regions, while our method exploits the statistical relations among different AUs from target AU labels, and the statistical AU label distribution is used as a constraint to supervise the learning process of deep network and AU classifier. When comparing with shallow models, like LSVM and JPML, our method also achieves better performance. The average F1 Score of ours is 14% and 3% higher than LSVM and JPML respectively. Though JPML defines AU relations through dataset s-

Table 1. F1 Score on the BP4D database.

AU	Our	without label constraint	DRML [25]	AlexNet [25]	ConvNet [25]	LCN [25]	LSVM [25]	JPML [25]
1	0.381	0.287	0.364	0.270	0.404	**0.450**	0.232	0.326
2	0.166	0.187	0.418	0.255	**0.461**	0.412	0.228	0.256
4	0.418	0.310	**0.430**	0.319	0.428	0.423	0.231	0.374
6	**0.741**	0.737	0.550	0.514	0.518	0.586	0.272	0.423
7	0.620	0.666	**0.670**	0.554	0.543	0.528	0.471	0.505
10	0.739	0.679	0.663	0.528	0.540	0.540	**0.772**	0.722
12	**0.792**	0.778	0.658	0.490	0.610	0.547	0.637	0.741
14	0.588	0.470	0.541	0.517	0.567	0.599	0.643	**0.657**
15	0.246	0.202	0.332	0.255	**0.441**	0.361	0.184	0.381
17	**0.564**	0.503	0.480	0.414	0.383	0.466	0.330	0.400
23	0.261	0.307	0.317	0.261	**0.418**	0.332	0.194	0.304
24	0.376	0.228	0.300	0.235	0.328	0.353	0.207	**0.423**
Avg.	**0.491**	0.446	0.483	0.384	0.470	0.466	0.353	0.459

Table 2. F1 Score on the DISFA database.

AU	Our	without label constraint	GPDE [3]	MC-LVM [4]	HRBM [4]	l_p-MTMKL [4]
1	0.574	**0.608**	-	0.586	0.397	0.422
2	0.570	0.489	-	**0.630**	0.559	0.458
4	0.658	0.665	0.656	**0.729**	0.616	0.472
6	**0.684**	0.594	0.536	0.523	0.540	0.628
9	0.334	0.337	**0.471**	-	-	-
12	0.791	0.797	0.600	**0.847**	0.792	0.763
15	0.472	0.363	-	**0.494**	0.387	0.345
17	0.447	0.201	-	**0.486**	0.388	0.414
25	0.636	0.625	**0.800**	-	-	-
26	**0.675**	0.643	0.571	-	-	-
Avg.	0.584	0.532	0.606	**0.614**	0.526	0.500



(a) raw image pixels on DISFA (b) deep features on DISFA (c) raw image pixels on BP4D (d) deep features on BP4D

Figure 3. A t-SNE embedding of raw image pixels and learnt deep features in term of subjects on the BP4D and DISFA database. DISFA: colored in AU25; BP4D: colored in AU12. Each text represents one subject ID and is placed at the center of its own frames. The clustering effect reveals that face images retain individual differences; the learnt deep representation reduces such influence.

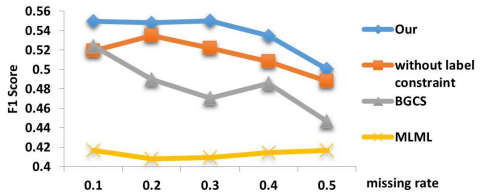
tatistics, it uses manually-crafted feature (i.e., SIFT). Our method learns the feature representations using a deep neural network. What's more, in our model, the AU relations are not only used to train the AU classifiers, but are also adopted to fine-tune the feature learning process. Linear SVM (LSVM) is a data-driven method, and it ignores the relations among different AUs. The superiority of our method over the shallow models demonstrates the effectiveness of deep framework in learning effective feature representations, and also proves the importance of AU relations in AU recognition. What's more, the method of without AU label distribution constraint is also superior to LSVM, further demonstrating the effectiveness of deep model in extracting features.

The results of the DISFA database are listed in Table 2. Since the number of samples and AUs on the DISFA database used in each related work are not the same, it is hard to compare our work with those in a totally fair way. So the comparison on this database is only for reference. In terms of the common AUs, our method achieves an average F1 Score of 0.630, which is 2% higher than GPDE. For the AUs used in [4], the result of ours is 0.599, which is better than the state-of-the-art approaches in most cases. In more detail, the average F1 Score of our method is about 7% and 10% higher than HRBM and l_p-MTMKL respectively.

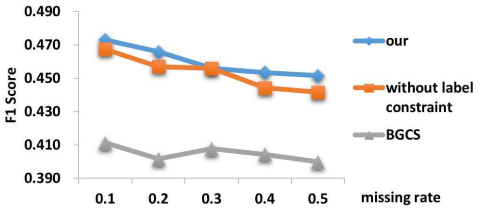
5.2.2 AU recognition on partially labeled data

As discussed in Section 4, our proposed method can learn AU classifiers from partially labeled data. To validate the effectiveness of our method in semi-supervise learning, we perform experiment with partially labeled data with/without the help of AU label distribution constraint. As mentioned in Section 2, there are several recent works considering AU classification under partially AU annotations. [17] required expression labels as hidden knowledge to complement the missing AU labels, and [13] also required a large-scale expression-labeled images. Thus we only compare our work with MLML [20] and BGCS [15], which train AU classifier with missing labels without extra label efforts. For a fair comparison, we rerun their code using our data. Since MLML involves in calculating the similarities between two images, the size of similarity kernel function is proportional to the square of the number of samples, resulting in a memory problem. Therefore, on the BP4D database, we only compare our method with BGCS.

Figure 4(a) and Figure 4(b) depict the results on the DISFA database and the BP4D database respectively. As can be seen, in general, the AU recognition performance decreases with the increasing of missing rate. On both databases, our proposed method performs better than the method without AU label distribution constraint in most cases, demonstrating the effectiveness of AU label distribution constraint in regularizing the learning process of deep network and AU classifier. Compared with BGCS, our method outperforms in most cases, since the average F1 Scores of ours are higher than those of BGCS on both databases. What's more, compared with MLML, our method performs better as well. Both BGCS and MLML used hand-craft features, while we learn feature representation from deep framework, which is more effective. Furthermore, they either use label smoothness to fill in the missing values or marginalize over the missing values during inference. In our model, we enforce the output of unlabeled data to be consistent with AU label distribution learnt from ground-truth target labels. These leads to the superiority of our method in semi-supervise learning.



(a) DISFA



(b) BP4D

Figure 4. AU recognition results with partially labeled data under 5 different missing rates.

6. Conclusions

Traditional machine learning approaches to AU recognition require large amounts of labeled data. This requirement is often unrealistic, since AU annotation is a time consuming task. Thus, in this work, we propose a novel model to tackle the problem of AU recognition under incomplete AU annotations. Due to the facial anatomy, there exists spatial patterns in multiple AUs. These relations can be used to help AU recognition. Hence the AU label relations described in a probability form, namely AU label distribution, are incorporated to learn the AU classifier. First, a deep neural network is learnt for feature representation from a large-scale facial images. Then, a restricted Boltzmann machine is trained to learn the AU label distribution from available AU labels. Finally, we train the AU classifier by maximizing the log likelihood of the mapping functions of AUs with regard to the learnt label distribution and minimizing the error between predicted AUs and ground-truth AUs from partially labeled data. The experimental results under complete AU annotations prove that AU label distribution constraint can achieve a better performance on AU recognition and demonstrate the learnt features from the deep framework can diminish the individual influence. The experimental results with partially labeled data demonstrate the superiority of our method in semi-supervise learning of AU recognition.

Acknowledgment

This work has been supported by the National Science Foundation of China (Grant No. 61473270, 61228304, 61175037), and the project from Anhui Science and Technology Agency (1508085SMF223).

References

[1] W.-S. Chu, F. De la Torre, and J. F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016.

[2] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15):E1454, 2014.

[3] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Gaussian process domain experts for model adaptation in facial behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26, 2016.

[4] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3792–3800, 2015.

[5] A. Fischer and C. Igel. An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36, 2012.

[6] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 609–615. IEEE, 2015.

[7] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–5. IEEE, 2015.

[8] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[9] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[10] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[11] Y. Li, B. Wu, B. Ghanem, Y. Zhao, H. Yao, and Q. Ji. Facial action unit recognition under incomplete data based on multi-label learning with missing labels. *Pattern Recognition*, 60:890–900, 2016.

[12] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[13] A. Ruiz, J. Van de Weijer, and X. Binefa. From emotions to action units with hidden and semi-hidden-task learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3703–3711, 2015.

[14] A. S. K. Shizhong Han, Zibo Meng and Y. Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Neural Information Processing Systems (NIPS)*, pages 1–9, 2016.

[15] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[16] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence*, 29(10), 2007.

[17] S. Wang, Q. Gan, and Q. Ji. Expression-assisted facial action unit recognition under incomplete au annotation. *Pattern Recognition*, 61:78–91, 2017.

[18] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013.

[19] Z. Wang, S. Lyu, G. Schalk, and Q. Ji. Learning with target prior. In *Advances in Neural Information Processing Systems*, pages 2231–2239, 2012.

[20] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7):2279–2289, 2015.

[21] Y. Z. Yongqiang Li, Jixu Chen and Q. Ji. Data-free prior model for facial action unit recognition. *IEEE Transactions on Affective Computing*, 2013.

[22] X. Zhang and M. Mahoor. Simultaneous detection of multiple facial action units via hierarchical task structure learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1863–1868, Aug 2014.

[23] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[24] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[25] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[26] Y. Zhu, S. Wang, L. Yue, and Q. Ji. Multiple-facial action unit recognition by shared feature learning and semantic relation modeling. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1663–1668, Aug 2014.