Exploring Domain Knowledge for Facial Expression-Assisted Action Unit Activation Recognition

Shangfei Wang, Senior Member, IEEE, Guozhu Peng, and Qiang Ji, Fellow, IEEE

Abstract—Current works on facial action unit (AU) activation recognition typically include supervised training using AU-annotated training images. Compared to facial expression labeling, AU annotation is a time-consuming, expensive, and error-prone process. Domain knowledge refers to the strong probabilistic dependencies between facial expressions and AUs, as well as dependencies among AUs. To take advantage of this, we avoid the time-consuming process of AU annotation and introduce a new AU activation recognition method that learns AU classifiers from domain knowledge, and requires only expression-annotated facial images. Specifically, we first generate pseudo AU labels according to the probabilistic dependencies between expressions and AUs as well as correlations among AUs summarized from domain knowledge. Then, we propose to use a Restricted Boltzmann Machine to model AU label prior distribution from the generated pseudo AU data. After that, we train AU classifiers from expression-annotated facial images and the learned prior model by maximizing the log likelihood of AU classifiers with regard to the learned AU label prior. The proposed AU activation recognition can also be extended to semi-supervised learning scenarios with partially AU-annotated facial images. Experimental results on four benchmark databases demonstrate the effectiveness of the proposed approach in learning AU classifiers from domain knowledge.

Index Terms—AU activation recognition, domain knowledge, weakly supervised, expression-assisted, RBM

1 INTRODUCTION

With the development of user-centered human-computer interaction, automatic facial expression analysis has attracted increasing attention in the fields of human behavior research [1] and computer vision [2], [3], [4], [5], since facial expression is one of the most important channels for human-human emotion communication [6], [7]. Facial expressions can be presented by expression categories and facial action units (AUs). For expressions, anger, disgust, fear, happiness, sadness, and surprise (i.e., Ekman’s six universal expression categories) are frequently used in automatic facial expression analyses. However, in real scenarios, other distinct faces such as contemptibility, hatred, awe, pain and combinations of the basic expressions may also occur. Due to the diversity of human expressions, Ekman et al.’s facial action coding system (FACS) [8] is frequently used for automatic facial expression analyses. FACS can systematically categorize the physical expression of emotions through 44 main AUs, several head movement codes, eye movement codes, visibility codes, and gross behavior codes. Nearly all anatomically possible facial expression can be deconstructed into several specific AUs and their temporal segments.

Many annotated training images are required to recognize AUs and expressions. AU annotation is harder and more expensive than expression annotation, since expressions are global and easier to recognize, while AUs are local, subtle, and more difficult to identify. Furthermore, there may be several AUs for one image, but there are usually only a few expressions. Therefore, the AU labels should be provided by qualified FACS experts. In addition, AU labels are prone to error, since some AUs are subtle and difficult to annotate. While expressions are much easier to label accurately. It is easy to locate expression-labeled web images, but AU-labeled web images are far scarcer, providing additional evidence that AU annotation is more expensive and difficult than expression annotation.

Expression categories and facial action units must be closely related, since they both describe facial behaviors. Specifically, expression categories describe facial behaviors in a global way, while facial action units depict the local variations on a face. Furthermore, most people express an emotion using the same facial muscles. For example, Du et al. [9] found that 99 percent of the time, people show happiness with a smile by stretching their mouths and raising their cheeks. Emotional Facial Action Coding System (EMFACS) deconstructs expressions into several facial actions [10], e.g., AU6 and AU12 usually appear simultaneously on happy faces. Prkachin et al. [11] found that pain intensity can be interpreted from the combination of AU4, AU6, AU7, AU9,
AU10, and AU43. Fig. 1 shows the six basic expressions and their combinations of AUs, taken from the CK+ database. These expression-AU dependencies are consistent with domain knowledge observed in behavior research. For example, as shown on the CK+ database, more than 90 percent of happy faces consist of AU12 and AU25, and about 60 percent of fear faces include AU5. This is consistent with Du et al.’s [9] observations, as shown in Table 1. Such inherent expression-AU relations can be exploited to learn AU classifiers from images with expression labels but without AU labels.

Therefore, our paper aims to learn AU classifiers from images with complete expression labels but lack of AU annotation under the help of the domain knowledge of expression-AU dependencies as well as AU-AU dependencies. To the best of our knowledge, only one work learns AU classifiers without AU labels. Ruiz et al. [12] proposed Hidden-Task Learning (HTL) to learn AU classifiers from large-scale facial images labeled with six universal facial expressions through exploiting domain knowledge about the relation between expressions and AUs. They first sampled AU data based on the dependencies between six basic expressions and AUs summarized from [13] and [14], and then trained AU classifier from images and expression classifier from AUs with sampled AU data and the large-scale facial images. Unlike Ruiz et al.’s work, which learns AU classifiers from extra large-scale expression-labeled facial images and expression classifiers from AUs, we train AU classifiers from facial images with expression labels directly. Thus, we do not need another large-scale expression-annotated facial images database, and avoid any error propagated to AU classifiers by expression classifiers. Furthermore, instead of using only the dependencies between six basic expressions and AUs as Ruiz et al.’s work, we exploit both expression-dependent and expression-independent AU relations in our work. For expression-dependent AU relations, we consider six basic expressions and pain expression. Specifically, we first summarize expression-dependent AU probabilities and expression-independent AU probabilities from facial anatomy and behavior research as the domain knowledge. Then we translate the domain knowledge into pseudo AU data through sampling based on the summarized probabilities for each expression. After that, a RBM is used to capture the AU label distribution of each expression from pseudo AU data, and AU classifiers are learned by maximizing the log likelihood of AU classifiers with regard to the learned AU label prior from both facial images and label distributions. The framework of the proposed method is shown in Fig. 2.

2 RELATED WORK

Currently, the mainstream of AU activation recognition research classifies each AU independently or detects certain AU combinations. These methods either ignore the inherent dependencies among multiple AUs or cannot handle hundreds of variance in AU combinations. Only recently, several works exploited AU relations from AU labels or facial appearance for AU activation recognition. Probabilistic graphic models are adopted to capture AU dependencies from AU labels. For example, Tong et al. [15] adopted the structure and the conditional probabilities of the learned Bayesian Network (BN) to model the semantic relationships among AUs. Wang et al. [16] used the connections between the hidden units and visible units of a hierarchical Restricted Boltzmann Machine (RBM) to capture the global semantic relationships among AUs. These two studies exploit AU dependencies from ground truth AU labels to facilitate AU activation recognition. Unlike the above two works, Li et al. [17] proposed to learn AU dependencies from pseudo-data, which are generated from domain knowledge.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Prototypical (and variant AUs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>4, 7, 24 [10(26%), 17(52%), 23(29%)]</td>
</tr>
<tr>
<td>Disgust</td>
<td>9, 10, 17 [431%, 24(26%)]</td>
</tr>
<tr>
<td>Fear</td>
<td>1, 4, 20, 25 (25%5, 563%, 2633%)]</td>
</tr>
<tr>
<td>Happiness</td>
<td>12, 25 [65%]</td>
</tr>
<tr>
<td>Sadness</td>
<td>4, 15 [16%, 65%, 1126%, 1767%]</td>
</tr>
<tr>
<td>Surprise</td>
<td>1, 2, 25, 26 [566%]</td>
</tr>
<tr>
<td>pain</td>
<td>4(&gt; 50%), 6(&gt; 50%), 7(&gt; 50%), 9(&gt; 50%), 10(&gt; 50%), 43(&gt; 50%)</td>
</tr>
</tbody>
</table>

Fig. 1. Examples of six basic expressions and corresponding AU combinations on the CK+ database. Upper row: Image samples of six basic expressions. Lower row: AU distributions of corresponding facial expressions.
Instead of modeling AU relations from target labels only, several works on AU activation recognition leverage the AU dependencies inherent in both target labels and image features. For example, Zhu et al. [18] adopted multi-task learning to capture feature dependencies among AUs of each AU group, and a Bayesian network to model label relations. Zhang et al. [19] detected fixed groups of multiple AUs simultaneously by using multi-task multiple kernel learning. The fixed AU groups are determined by a hierarchical model based on AU co-occurrences existing in both AU labels and facial regions. Zhao et al. [20] proposed to select a sparse subset of facial patches, and learn multiple AU classifiers simultaneously under the constraints of group sparsity and local AU relations of positive correlation and negative competition. Instead of integrating predefined local label dependencies into classifiers in the original feature space or kernel space as in the above works, Eleftheriadis et al. [21] integrated AU label dependencies into latent space and classifier learning through the proposed multi-conditional latent variable model. Zhao et al. [22] proposed deep region and multi-label learning to identify discriminative facial regions and detect multiple AUs simultaneously.

All these works successfully exploit the dependencies inherent in multiple AUs to facilitate AU activation recognition. However, few of them consider the relations between AUs and expressions to help AU activation recognition, and all of them require complete AU annotations to train AU classifiers. Only recently, a few works turn to AU activation recognition from partially AU annotated samples. Song et al. [23] proposed a Bayesian graphical model to encode sparsity and co-occurrence structure of facial action units via compressed sensing and group-wise sparsity inducing priors. Their proposed method handles partially observed labels through marginalizing over the unobserved values. Wu et al. [24] explored the consistency between the predicted labels and the provided labels as well as the local smoothness among the label assignments for multi-label classification with missing labels (MLML). Instead of using the same features for all AU classes as Wu et al.’s work, Li et al. [25] extended the MLML method to discriminate each AU based on the most related features. They either take advantage of the attributes of the generative model to handling missing AU labels or adopt label consistency and smoothness as constraints to facilitate AU classifier learning from partial AU labels. All of these work require at least partial AU labels to learn AU classifiers.

To the best of our knowledge, so far only three works recognize AUs with the help of expressions. Wang et al. [26] proposed an expression-assisted AU activation recognition method under incomplete AU labeling. They constructed a Bayesian network to capture both the dependencies among AUs and the dependencies between AUs and expressions, and adopted Structure Expectation Maximization (SEM) to learn the structure and parameters of the Bayesian network when AU labels are missing. After training, the AUs of testing images are inferred by combining the measurements and the AU relations in the BN model. Their proposed method successfully uses expression labels as hidden knowledge to complement the missing AU labels. However, both expression labels and AU labels are required to learn AU classifiers, although AU labels may be partially missing.

Wang et al. [16] proposed an AU activation recognition method using expression as privileged information, which is only required during training. A 3-way RBM is used to capture the global dependencies between expressions and AUs. Their method requires both complete AU and expression labels during training. Ruiz et al. [12] proposed to learn AU classifiers from un-annotated facial images with the help of additional large scale facial images labeled with expressions but not AUs. Their method separately trains one classifier from AU to expression and another classifier from feature to AU. Domain knowledge about the relation between six basic expressions and AUs is employed to learn expression classifiers from AU classifiers first, and then AU classifiers from images can be learned through embedding the output of AU classifiers as the input of expression classifiers. Ruiz et al.’s method successfully leverages the domain knowledge of expression-AU relations to train AU classifiers from the training data when AU labels are limited or unavailable. However, their method only considers expression-dependent AU relations related to six basic expressions, and requires a large-scale expression-labeled images database. Furthermore, any error caused by expression classifiers will propagate to AU classifiers.

3 Problem Statement

The purpose of our work is to learn weakly supervised AU classifiers from images with expression labels but without AU labels. Considering the strong relations between AUs and expressions, the basic idea is to learn AU probabilistic distributions given expressions first, and then maximize the log likelihood of AU classifiers with regard to the learned AU distributions.

Let \( T = \{ \Gamma_1, \Gamma_2 \} \) denote the domain knowledge of expressions and AUs. It includes both expression dependent conditional probabilities of AUs \( \Gamma_1 \) and expression independent conditional probabilities of AUs \( \Gamma_2 \). To digitalize the domain knowledge, we first generate \( K \) pseudo AU samples \( Y = \{ Y_i \}_{i=1}^{N} \), where \( Y_i = \{ y_{i1}, y_{i2}, \ldots, y_{iK} \} \), according to \( T \) under each expression category. Then we learn the target prior, i.e., AU distributions, \( P_\theta(y|E) \) from the sampled AU data.

Let \( X = \{ (x^n, E^n) \}_{n=1}^{N} \) denote the training set, where \( N \) is the number of training samples, \( x^n \in \mathbb{R}^d \) represents the \( d \)-dimensional feature vector. \( E^n \in \{ 1, 2, \ldots, P \} \) represents the expression label and \( P \) is the number of expression categories. Our goal is to learn a classifier \( f : \mathbb{R}^d \rightarrow \{ 0,1 \}^L \) according to the follow equation:

\[
\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log P_\theta(f(x^n; \Theta)|E^n, T),
\]
where $\Phi$ are the parameters of AU distribution and $\Theta$ are the parameters of AU classifier.

The proposed method can be extended to semi-supervised learning method when AU labels are available for partial training samples. Let $X^m = \{ (x^m, y^m) \}_{m=1}^M$ denote the subset of fully AU labeled training data, where $M$ is the number of training samples, $x^m \in \mathbb{R}^d$ represents the $d$-dimensional feature vector. $y^m = \{ y^m_1, y^m_2, \ldots, y^m_n \}$ represents the $L$-dimensional ground truth AU labels. Then, the objective in Equation (1) can be extended

$$
\max_{\Theta, \Phi} \frac{1}{N} \sum_{n=1}^N \log P_\theta(f(x^n; \Theta)|E^n, T) - \frac{1}{M} \sum_{m=1}^M \mathcal{L}_\Theta(x^m, y^m), \quad (2)
$$

where $\mathcal{L}_\Theta$ is the loss function for the fully-labeled data and $\alpha$ is the trade-off rate between the two terms in Equation (2).

## 4 Proposed Method

The proposed method consists of four steps. First, we summarize the domain knowledge about expressions and AUs. Second, we generate pseudo AU data from domain knowledge for each expression category. Then, we learn AU prior from the generated pseudo AU data. Specifically, we use RBM to model AU distribution since RBM can capture high-order dependencies among visible nodes, through the connections between visible nodes and hidden nodes. Lastly, we train AU classifiers from expression-annotated facial images and the learned AU label distribution. The object function is to maximize the log likelihood of AU classifiers with regard to the learned AU label prior. We can also extend the proposed method to use in semi-supervised scenarios with partially labeled data.

### 4.1 Summary of Domain Knowledge on AUs and Expressions

In this section, we summarize the domain knowledge of expressions and AUs [27], i.e., expression-dependent probability of AUs $Y_1$ and expression-independent probability of AUs $T_2$. The former can be categorized into two types: the marginal probability of a single AU given an expression, and the conditional probability of one AU under another AU and expression. The later is the conditional probability of one AU under another AU, including both co-existent and mutually exclusive relations.

In this paper, we consider pain and six basic expressions. For expression-dependent AU relations, we first consider the domain knowledge about the six basic expressions and AUs. For marginal probability of a single AU given an expression, we adopt the observations from [9], as shown in the first six rows of Table 1. The expression-dependent marginal probability of AUs which are not followed by parentheses, is larger than 70 percent. The expression-dependent marginal probability of AUs, which are not listed, is less than 20 percent. The percentage in parentheses following the AU is the marginal probability of a single AU, given the expression. For example, from the first row of Table 1, we know: $P(AU4 | \text{anger}) \geq 70\%$, $P(AU17 | \text{anger}) = 52\%$ and $P(AU1 | \text{anger}) < 20\%$. These probabilities can provide weak supervisory information for classifier’s training.

For the conditional probability of one AU under another AU and expression, we adopt the domain knowledge from the Emotion Facial Action Coding System [10] as shown in Table 2. Table 2 lists AU combinations, i.e., co-existent relations among AUs, for each expression. For example, though AU6 and AU15 don’t correlate to each other according to facial anatomy, they almost always appear simultaneously during sadness, i.e., $P(AU6 \& AU15, \text{sadness})$ is larger than 70 percent, which is the probability of a chance event.

In addition to the domain knowledge about six basic expressions, we also have domain knowledge about pain expression according to Prkachin et al.’s work [11]. Prkachin et al. [11]’s observations indicate that four actions, i.e., brow lowering (AU4), orbital tightening (AU6 and AU7), levator contraction(AU9 and AU10) and eye closure(AU43), carry the bulk of information about pain. Afterwards, Prkachin et al. [28] explicitly found that pain is strongly associated with these six AUs as the defined Prkachin and Solomon pain intensity (PSPI) shown

$$
PSPI = AU4 + (AU6orAU7) + (AU9orAU10) + AU43. \quad (3)
$$

Equation (3) suggests that these six AUs play a significant role in pain expression. Therefore, we think the occurrence probability of each of these AUs should be higher than chance, i.e., 50 percent for pain frames, as shown in Table 1. For example, $P(AU6 \& \text{pain}) > 50\%$.

The expression-independent AU probabilities are mainly caused by the muscular structure of human face, and universal for all expressions. They include both co-existent and mutually exclusive relations. For example, *lip tightener (AU23)* and *lip pressor (AU24)* are both related to the muscle group *orbicularis oris*. Most people cannot make a facial movement of AU23 without AU24, and vice versa. It means both $P(AU23 \& AU24)$ and $P(AU24 \& AU23)$ are very large. This is a kind of co-existent relation. *chin raiser (AU17)* rarely appears with *chin depressor (AU17)*. The former is produced by the muscle group *zygomaticus major*, and the latter is produced by the muscle group *mentalis*. Therefore, $P(AU17 \& AU12)$ and $P(AU12 \& AU17)$ are very small. This is referred as a mutually exclusive relation. Table 3 lists expression-independent AU relations summarized from [8].

<table>
<thead>
<tr>
<th>Expression-Related AU Combinations from EMFACS [10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Anger</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td>Disgust</td>
</tr>
<tr>
<td>Happiness</td>
</tr>
<tr>
<td>Sadness</td>
</tr>
<tr>
<td>Surprise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expression-Independent AU Relations [8], [17]</th>
<th>coexistent</th>
<th>mutually exclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU1—AU2—AU5</td>
<td>AU12—AU15, AU12—AU17</td>
<td></td>
</tr>
<tr>
<td>AU4—AU7—AU9</td>
<td>AU2—AU6, AU2—AU7, AU2—AU9</td>
<td></td>
</tr>
<tr>
<td>AU15—AU17—AU24</td>
<td>AU15—AU25, AU17—AU25</td>
<td></td>
</tr>
<tr>
<td>AU23—AU24</td>
<td>AU23—AU25, AU24—AU25</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2**

**TABLE 3**
and [17]. For example, from the first row of Table 3, we know: AU1 and AU2 are coexistent, AU12 and AU15 are mutually exclusive.

For coexistent AU relations, whether it’s expression-dependent or expression-independent, the conditional probability of one AU given another AU must be larger than chance, i.e., 50 percent. For mutually exclusive AU relations, following Du et al.’s work, we set the probability of exclusive relations as less than 20 percent, since Du et al. did not list the AUs whose expression-dependent marginal probability less than 20 percent in Table 1. For example, P(AU6 | AU12, happiness) > 50%, P(AU1 | AU2) > 50%, and P(AU12 | AU15) < 20%.

### 4.2 Pseudo AU Label Generation from Domain Knowledge

After summarizing both expression-dependent and expression-independent AU probabilities from domain knowledge, we generate pseudo AU labels based on the summarized domain knowledge following Peng et al.’s work [27]. Since some AU probabilities are expression-dependent, pseudo AU labels are generated for each expression. First, we decide the sampling order of AUs according to the ascending order of their expression-dependent marginal probability, found in Table 1. Then, for the first AU, we generate pseudo AU labels according to its expression-dependent marginal probability. For the rest of the AUs, if they do not have relations to already-generated AUs, we generate pseudo AU labels according to the expression-dependent marginal probability listed in Table 1. Otherwise, we generate pseudo AU labels according to the coexistence relations and mutually exclusive relations with existing AUs, as listed in Tables 2 and 3 respectively. Specifically, if AU_s and AU_j are co-existent, we draw P(AU_j | AU_s) from the uniform distribution U(0.5, 1) and its mean value is 0.75. Therefore, we set P(AU_j = 1 | AU_s = 1) = 0.75 + ε, ε is drawn from U(-0.25, 0.25). If they are mutually exclusive, we draw P(AU_j = 1 | AU_s = 1) from U(0, 0.2), and its mean is 0.10. Therefore, we set P(AU_j = 1 | AU_s = 1) = 0.10 + ε, ε is drawn from U(-0.1, 0.1). Detailed sampling algorithms are shown in Algorithm 1. In our work, we sample 5,000 pseudo AU samples under each expression, i.e., K = 5,000 in Algorithm 1.

### 4.3 Learning AU Prior from Generated AU Data

Based on the generated pseudo AU data, we model the AU prior distribution through the Restricted Boltzmann Machine [29]. The RBM is a generative model, which can model distribution of input variables based on latent variables. In our work, we use a Bernoulli-Bernoulli RBM consisting of two layers: visible layer y and hidden layer h. The visible nodes represent AU labels. A graphical depiction of the RBM is shown in Fig. 3.

![Fig. 3. The structure of AU prior model.](image)

The energy function of RBM is defined as

$$E(h, y; \Phi) = - \sum_{i=1}^{L} W_{yi} y_i h_i - \sum_{i=1}^{L} b_i y_i - \sum_{j=1}^{q} a_j h_j, \quad (4)$$

where $\Phi = \{W, b, a\}$ are parameters of RBM, $b_i$ represents bias of $i$th AU, $a_j$ represents the bias of $j$th hidden unit, and $W_{ij}$ represents the weight between $i$th AU and $j$th hidden unit. $L$ and $q$ represent the number of AUs and hidden units respectively.

### Algorithm 1. The Sampling of Pseudo AU Data. [27]

**Input:** The domain knowledge about expressions and AUs listed in Tables 1, 2 and 3.

**Output:** The pseudo AU samples $Y = \{Y_i\}_i$.

**for** expression $E_p (p = 1, 2, \ldots, P)$ **do**

- sort AUs according to the probability of AUs from small to large under expression $E_p$

  **repeat**

  - generate sample $Y_i = \{y'_1, y'_2, \ldots, y'_L\}$:
    - sample the first AU with $P(AU_1 | E)$ from Table 1
    - for $j = 2, 3, \ldots, L$ do
      - sample the $j$th component of $Y_i$, $y'_j$, i.e., the $j$th AU of sample $Y_i$, $AU_j$:
        - if any AU_j related to AU_j is already generated then
          - sample AU_j with $P(AU_j | AU_s, E)$ from Table 2 or $P(AU_j | AU_s)$ from Table 3.
        - else
          - sample AU_j with $P(AU_j | E)$ from Table 1
      - end if
    - end for

  **until** we have already got $K$ samples

**end for**

We can get the joint prior probability $P(y)$ through marginalizing the hidden nodes by

$$P(y) = \frac{1}{Z} \sum_{h} e^{-E(h, y; \Phi)} = \frac{1}{Z} \prod_{i=1}^{L} e^{h_i y_i} \prod_{j=1}^{q} (1 + e^{a_j + \sum_{i=1}^{L} W_{ij} y_i}), \quad (5)$$

where $Z = \sum_{y} \sum_{a} \exp(-E(h, y; \Phi))$ is a normalizing constant called the partition function.

From Equation (5), we can find that a larger weight results in a higher probability of $P(y)$. In contrast, a smaller weight represents a lower probability of $P(y)$. In other words, the weight between hidden units and AU labels measures the presence and absence of AUs.

We employ a Maximum Likelihood Estimation (MLE) method to learn the parameters $\Phi$ as shown in Equation (6). Due to the complexity of calculating the normalizing factor, we use the Contrastive Divergence (CD) [30] to estimate parameters $\Phi$. The gradient of $\Phi$ is shown in Equation (7).

$$\Phi^* = \arg \max_{\Phi} \log P(y; \Phi) \quad (6)$$

$$\frac{\partial \log P(y)}{\partial \Phi} = \langle \frac{\partial E}{\partial \Phi} \rangle_{P(h|y)} - \langle \frac{\partial E}{\partial \Phi} \rangle_{P(h|y)} \quad (7)$$
4.4 AU Classifier Training from Expression-Annotate Facial Images and AU Prior Model

As stated in Section 3, let $X = \{(x^n, E^n)\}_{n=1}^N$ denote the training set, where $N$ is the number of training samples, $x^n \in \mathbb{R}^d$ represents the $d$-dimensional image feature. $E^n \in \{1, 2, \ldots, P\}$ is the expression label and $P$ is the number of expression classes. Our goal is to learn an AU classifier $f : \mathbb{R}^d \rightarrow \{0, 1\}^L$ from facial image samples with expression labels only, where $L$ represents the dimension of AU vector.

For each expression category, we train a joint prior distribution of AUs $P(y|E)$ as an AU label constraint according to Section 4.3. The objective of our method is to maximize the log likelihood of AU classifier $f(x; \Theta)$ as shown

$$
\max_{\Theta} \frac{1}{N} \sum_{n=1}^N \log P(f(x^n; \Theta)|E^n) - \beta R(\Theta),
$$

(8)

where $\Theta$ is the parameter of the AU classifier, $R(\Theta)$ is the regularized term and $\beta$ is the hyper parameter determined by cross validation.

According to Equation (5), we could write the log $P(y)$ as Equation (9), and the derivation of $\log P(y)$ over $y$ is given by Equation (10), where $\sigma(s) = 1/(1 + \exp(-s))$ is a sigmoid function

$$
\log P(y) = \sum_{i=1}^L y_i b_i + \frac{1}{\alpha(n)} \sum_j \log(1 + \exp(a_j + \sum_{i=1}^L y_i W_{ij})) - \log Z
$$

(9)

$$
\frac{\partial \log P(y)}{\partial y} = b + \frac{1}{\alpha(n)} \sum_j \sigma(a_j + \sum_{i=1}^L y_i W_{ij}) W_{ij}
$$

(10)

In this paper, we use a linear function with a sigmoid output layer as the AU classifier as shown

$$
f(x; \Theta) = \sigma(V \ast x + g).
$$

(11)

So, $\Theta = \{V, g\}$, and the derivatives of $f(x; \Theta)$ over $V$ and $g$ are shown

$$
\nabla_V f(x) = f(x) \ast (1 - f(x)) \ast x
$$

$$
\nabla_g f(x) = f(x) \cdot (1 - f(x)).
$$

(12)

For regularized term, we use standard L2-regularization $\frac{1}{2}||\Theta||^2$, so $\nabla R(\Theta) = \Theta$. Thus the parameter $\Theta = \{V, g\}$ can be updated using gradient ascent as

$$
\Theta^{(t+1)} = \Theta^{(t)} + \gamma \left[ \frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(f(x^n; \Theta)|E^n)}{\partial f(x^n; \Theta)} \frac{\partial f(x^n; \Theta)}{\partial \Theta} - \beta \Theta \right],
$$

(13)

where $t$ and $y^{(t)}$ represent the $t$th iteration and the learning rate of the $t$th iteration respectively.

4.5 Extension to Partially AU-Labeled Data

As stated in Section 3, when given partially labeled data, i.e., parts of $X$ are also labeled with AU labels, the proposed method can be extended to learn in a semi-supervised manner. Let $X = \{(x^m, y^m)\}_{m=1}^M (M \leq N)$, where $y^m = \{y_{1m}, y_{2m}, \ldots, y_{Lm}\} \in \{0, 1\}^L$, and $L$ is the number of AU labels. The updated objective is to solve the optimization problem defined in Equation (14), which differs from Equation (8) in that it includes an additional term minimizing the error between predicted AUs and the ground truth AUs

$$
\min_{\Theta} \alpha \frac{1}{N} \sum_{n=1}^N \log P(f(x^n; \Theta)|E^n) + \frac{1}{M} \sum_{m=1}^M \mathcal{L}_\Theta(x^m, y^m) + \beta R(\Theta).
$$

(14)

In this equation, the first term is the log likelihood of the AU classifier with regard to the learned target prior model for all training data. The second term $\mathcal{L}_\Theta$ defined in Equation (15) represents the cross-entropy loss function over the partially labeled data set $X^\alpha$, and the parameter $\alpha \in [0, 1]$ is a trade-off between the two terms

$$
\mathcal{L}_\Theta(x^n, y^n) = - [y^n \log f(x^n; \Theta) + (1 - y^n) \log (1 - f(x^n; \Theta))].
$$

(15)

Specifically, from Equation (14), we note that when $\alpha = 1$, the minimization problem is equivalent to learning in an un-supervised way as defined in Equation (8). Conversely, when $\alpha = 0$, it becomes a traditional supervised learning and does not consider the prior relations among multiple labels. In all other instances, the parameters are learned by maximizing the log likelihood of the AU classifier over the learned AU label distribution of all the samples, and minimizing the loss function of the partially annotated data. Similarly, a gradient ascent method is used to learn the parameters. The gradient of $\ell_\Theta$ over the parameters $\Theta$ is computed as

$$
\nabla_\Theta \mathcal{L}_\Theta(x^n, y^n) = - [y^n \cdot (f(x^n; \Theta))^{-1} - (1 - y^n) \cdot (1 - f(x^n; \Theta))^{-1}] \frac{\partial f(x^n, \Theta)}{\partial \Theta},
$$

(16)

where $\frac{\partial f(x^n, \Theta)}{\partial \Theta}$ can be calculated by Equation (12).

5 EXPERIMENTS

5.1 Experimental Conditions

Two posed and two spontaneous databases are used to evaluate the proposed weakly supervised AU activation recognition method: the Extended Cohn-Kanade database (CK+) [31], the MMI database [32], the Denver Intensity of Spontaneous Facial Action (DISFA) database [33], and the UNBC-McMaster Shoulder Pain Expression Archive database [34].

The CK+ database contains 593 posed expression sequences from 123 subjects, starting from the onset frame and ending with the apex frame. Among them, 309 sequences of 106 subjects are annotated with six basic expressions and AUs. We consider 12 AUs (1, 2, 4, 5, 6, 7, 9, 12, 17, 23, 24, and 25) whose occurrence frequencies are greater than 10 percent.

The MMI database contains 2,900 videos from 75 subjects. Among them, 171 sequences from 27 subjects are annotated with six basic expressions and AUs. In total, 13 AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 17, 23, 25, and 26) whose frequency of occurrence is larger than 10 percent of all samples are considered.

The DISFA database consists of facial expression videos from 27 subjects as they watch YouTube videos. Each frame is annotated with intensities of 12 AUs (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, and 26). According to AU intensity, 482 apex
frames are chosen. And we treat each AU with intensity larger than zero as active. Furthermore, the expression label of each frame is annotated using EMFACS. Similar to the CK+ database, 9 AUs (1, 2, 4, 6, 12, 17, 20, 25, and 26) with occurrence frequencies greater than 10 percent are considered.

The UNBC database consists of 200 facial videos from 25 shoulder pain suffered patients. Each frame is coded with PSPI. In this paper, we define the frames that have a PSPI of 5 or higher as “pain” and those which have a PSPI of 0 as “no pain”. We select all pain and no pain frames (7,319 frames in total) from 30 sequences of 17 subjects where pain frames exist. The six AUs (4, 6, 7, 9, 10, and 43) listed in Equation (3) are considered, since we only have the domain knowledge about these six AUs and pain expression.

We adopted facial feature points as features. For the CK+ database, the DISFA database, and the UNBC database, we used the feature points provided by the database constructors. For the MMI database, we extracted feature points with IntraFace [35]. We normalized feature points, so that the eye centers fall on the given positions for all images based on affine transformation. The model selection was adopted to determine hyper parameters. Specifically, we use a five fold cross-validation strategy. In each fold, we extracted quarter of the training set i.e., 20 percent of the whole database, as the validation set according to the subject. The number of hidden units and all of the weight coefficients were determined by grid search and the parameters which achieve best performance on validate set are used. The average F1-measure is used as the performance metric.

We conduct AU activation recognition experiments using facial images with expression labels only (i.e., weakly supervised learning) and facial images with expression labels and partial AU labels (i.e., semi-supervised learning). For both weakly supervised learning and semi-supervised learning scenarios, we conduct within-database experiments via five-fold subject-independent cross validation and cross-database experiments. For weakly supervised learning scenarios, we compare our work with state-of-the-art work, i.e., HTL [12], the only work that recognizes AUs without AU-labeled images. We re-conduct the experiments on the CK+, DISFA and UNBC database and conduct the experiment on the MMI database with the implementation of HTL method since Ruiz et al. [12] used a different experimental strategy, i.e., leave-one-subject strategy, and do not provide experimental results on the MMI database. Therefore, the used images and the recognized AUs of HTL are the same as ours, except for AU43 on the UNBC database, since there is no probability of AU43 in the domain knowledge used in Ruiz et al.’s work.

For semi-supervised learning scenarios, we randomly miss AU labels according to certain probabilities: 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent, and conduct AU activation recognition experiments for each missing rate five times. We compare our work with four state-of-the-art works, i.e., BGCS [23], MLML [24], BN [26] and SHTL [12]. The experimental conditions of these semi-supervised works are different from ours. Specifically, Song et al. [23] only conducted the semi-supervised experiments under 50 percent missing labels; Wu et al. [24] conducted semi-supervised experiments with four missing rates (20, 40, 60, and 80 percent) and adopted different performance metrics; Instead of missing all AUs of one image as we do, Wang et al. [26] adopted different missing strategy by missing one certain AU with a specific proportion; Ruiz et al. [12] conducted the semi-supervised experiments under 50 percent missing labels. Furthermore, none of them conducted experiments on the MMI database. Therefore, we re-conduct the semi-supervised experiments with the same experimental conditions.

5.2 Experimental Results and Analyses

5.2.1 Evaluation of Learned AU Prior

As discussed in Section 4.3, each latent unit in the RBM model can capture a specific pattern, which is measured by the weights between the latent units and AU labels. Larger weight indicates a higher probability of occurrence, and smaller weight denotes a higher probability of absence. To validate the effectiveness of our method for learning the AU prior distribution from domain knowledge and to graphically illustrate the global dependencies modeled by the prior model, we list several examples in Fig. 4. Figs. 4a and 4b depict patterns captured by two latent units in the prior model for fear expression on the CK+ database. From Fig. 4a, we find that AU4 and AU25 appear with a high probability, and in Fig. 4b, AU1, AU2 and AU5 occur with a high probability. These patterns are consistent with the relations described in Table 1. Likewise, Fig. 4c presents a pattern captured by a hidden unit of the learned prior model for anger expression. It shows that AU4, AU17, AU23 and AU24 appear with a high probability, which is in accordance with the description in Table 1. These examples strongly demonstrate that generating pseudo AU data according to domain knowledge is reasonable, and that the learned AU relations from the generated pseudo AU data are exactly consistent with the domain knowledge.

Fig. 4. Semantic relations between expressions and AUs captured by two latent units of RBM under fear expression and one latent unit of RBM under anger expression on the CK+ database. X-axis: AU index. Y-axis: The weight between h and y. Larger weight indicates high probability of occurrence, while smaller weight indicates high probability of absence.
5.2.2 Experimental Results of Weakly Supervised Learning

The within-database experimental results of AU activation recognition from images with expression labels only are listed in Table 4. From this table, we find that the proposed method performs better than HTL [12] on four databases, with a higher average F1-measure and higher F1-measures for most AUs. On the CK+ database, the average F1-measure of the proposed method is 0.7050, which is 50.06 percent higher than HTL, and for specific AUs, the proposed method achieves better performances on 11 out of 12 AUs; this improvement is more than 50 percent on AU1, AU4, AU9, AU12, AU17, AU23, and AU24. Compared with HTL, the proposed method underperforms in AU6. From the ground truth labels on the CK+ database, we find that about 70 percent of AU6 labels appear with happiness, and the frequency of AU6 in all samples with happiness is 95.65 percent. The domain knowledge we used in Table 1 is $P(AU6 \mid$ happiness) = 51%, and the domain knowledge Ruiz et al. used is summarized from [13] and [14], where $P(AU6 \mid$ happiness) = 94%. Therefore, the larger bias between our domain knowledge and the ground truth AU6 labels on the CK+ database causes the worse performance of the proposed method compared to HTL. We believe our summarized domain knowledge is more general and compete than Ruiz et al.’s, since our method outperforms HTL in most AUs.

On the MMI database, the average F1-measure of the proposed method is 0.5161, achieving 19.74 percent improvement over HTL. For specific AUs, the F1-measures of the proposed method are higher than HTL on 10 out of 13 AUs, and the improvements on AU1, AU4, AU7, and AU17 are more than 50 percent. On the DISFA database, the average F1-measure of the proposed method is 0.4236, which is 14.27 percent higher than HTL, and for specific AUs, the proposed method achieves better performances on 5 out of 9 AUs, including an improvement of more than 50 percent on AU1, and AU2. On the UNBC database, the average F1-measure of the proposed method is 0.3510, which is 49.23 percent higher than HTL. For specific AUs, the F1-measures of the proposed method are higher than HTL on all common AUs, and the improvements on AU4, AU7, and AU9 are more than 50 percent. These results strongly demonstrate that the proposed method works well for both posed and spontaneous expression recognition. The better performance on the UNBC database suggests that our method is not limited to basic emotion settings. As long as there is domain knowledge about the considered expressions, our method can perform well.

Both the proposed method and HTL sample pseudo AU data; rather than only modeling dependencies between expressions and AUs, we also capture global dependencies among AUs for each expression category from a RBM model. Particularly, there are strong expression-independent relations in some AU groups, such as AU1/AU2, AU1/AU5 and AU17/AU25. The correct discrimination of one AU can be helpful if it is coexistent to another, like AU1 and AU2. From Figs. 4a, 4b and 4c, we can see that the weight between AU1 and the hidden unit, as well as the weight between AU2 and the hidden unit are positive or negative simultaneously in three patterns, which indicates we successfully capture this dependency in our model. From Table 4, we find that the performances of our method on AU1 and AU2 are similar, but the performances of HTL on these two AUs differs greatly, which demonstrates that this dependency contributes on the good performances on these two AUs. Furthermore, HTL trains both AU classifier from images and expression classifier from AUs. Any error caused by expression classifiers may propagate to the AU classifiers. While the proposed method learns AU classifiers directly by leveraging both expression-dependent AU relations and expression-independent AU relations, thus it avoids error propagation and leads to superior performance of AU activation recognition.

For cross-database experiments, we compare the proposed method to HTL [12], and the results are listed in Tables 5 and 6. We find that the cross-database experimental results of the proposed method are poorer than the within-database experimental results in most cases. This is reasonable as there are often biases among databases. When compared to HTL, we observe the following. First, our method obtains a higher average F1-measure and higher F1-measures of most AUs in all experiments in Table 5 and the first three experiments in Table 6. Specifically, for the experiments that test on the UNBC database, when training on the CK+ database, the improvement is 27.24 percent; when training on the MMI database, the improvements is 5.72 percent; and when training on the DISFA database, the improvements is 5.91 percent. The UNBC is not a database of basic emotion settings. As a result, it strongly demonstrate that the proposed method works well for both posed and spontaneous expression recognition. Our improved results demonstrate that the proposed method successfully leverages more complete domain knowledge for better generalization ability. Second, for experiments that train on the UNBC database, the performances of the proposed method are worse than HTL. This is because HTL trains on another large facial image database with the same expression setting as the testing set, while the proposed method only trains on the UNBC database with pain expression. Thus when testing...
The within-database experimental results of semi-supervised AU activation recognition are illustrated in Fig. 5. From Fig. 5, we find that the performance of the proposed method decreases as the missing rate increases on all four databases. This is reasonable, since more AU labels could provide more supervisory information for AU classifier’s training.

For both weakly supervised and semi-supervised learning, the performances on the CK+ database are better than those on the MMI and the DISFA databases. The CK+ database is a posed expression database, while the MMI database and the DISFA database are spontaneous expression databases. It further proofs that it is more challenging to recognize spontaneous expressions than posed expressions.

The performances on the CK+, the MMI, and the DISFA databases are better than performances on the UNBC database. The CK+, the MMI and the DISFA databases include facial images with six basic expressions, while the UNBC database only consists of pain expression. The relations between AUs and six basic expressions have been studied more thoroughly than the relations between AUs and pain expression. Therefore, the summarized domain knowledge of six basic expressions provides more detailed and concrete expression-dependent AU probabilities than the summarized domain knowledge of pain expression. This more thorough domain knowledge provides better supervision for AU activation recognition.

Compared to four state-of-the-art semi-supervised AU activation recognition methods, we find that the proposed method performs best in most cases, demonstrating its superiority in handling missing AU labels for AU activation recognition. MLML leverages the label consistency and label smoothness to handle missing labels. BGCS handles partially observed labels by marginalizing over the unobserved values. BN adopts expression labels to complement missing AUs by the captured relationships between facial expressions and AUs. All above methods learn AU relations or AU-expression relations from partially available ground truth labels to handle missing AU labels. However, when the available ground truth labels are scarce, as in the scenarios with higher missing rates, these methods may not be able to learn AU relations well. For example, when the missing rate reaches 70 percent, the performance of MLML on the CK+, the MMI and the DISFA databases declines sharply, and when the missing rate reaches 90 percent, the performance of BN on the UNBC database also declines sharply. That doesn’t happen with our method, because we learn AU dependencies from domain knowledge, which doesn’t depend on the amount of ground truth AU labels. The AU relations coded in domain knowledge are more general than those embedded in ground truth AU labels, and lead to better performance. Both BGCS and BN use the Bayesian network to capture AU relations and BN also considers the assistance of expression labels, but Bayesian network can be applied to larger databases.

The performances on the CK+, the MMI, and the DISFA databases are better than performances on the UNBC database. The CK+, the MMI and the DISFA databases include facial images with six basic expressions, while the UNBC database only consists of pain expression. The relations between AUs and six basic expressions have been studied more thoroughly than the relations between AUs and pain expression. Therefore, the summarized domain knowledge of six basic expressions provides more detailed and concrete expression-dependent AU probabilities than the summarized domain knowledge of pain expression. This more thorough domain knowledge provides better supervision for AU activation recognition.

Compared to four state-of-the-art semi-supervised AU activation recognition methods, we find that the proposed method performs best in most cases, demonstrating its superiority in handling missing AU labels for AU activation recognition. MLML leverages the label consistency and label smoothness to handle missing labels. BGCS handles partially observed labels by marginalizing over the unobserved values. BN adopts expression labels to complement missing AUs by the captured relationships between facial expressions and AUs. All above methods learn AU relations or AU-expression relations from partially available ground truth labels to handle missing AU labels. However, when the available ground truth labels are scarce, as in the scenarios with higher missing rates, these methods may not be able to learn AU relations well. For example, when the missing rate reaches 70 percent, the performance of MLML on the CK+, the MMI and the DISFA databases declines sharply, and when the missing rate reaches 90 percent, the performance of BN on the UNBC database also declines sharply. That doesn’t happen with our method, because we learn AU dependencies from domain knowledge, which doesn’t depend on the amount of ground truth AU labels. The AU relations coded in domain knowledge are more general than those embedded in ground truth AU labels, and lead to better performance. Both BGCS and BN use the Bayesian network to capture AU relations and BN also considers the assistance of expression labels, but Bayesian network can be applied to larger databases.

TABLE 5

<table>
<thead>
<tr>
<th>AU</th>
<th>From CK+ to MMI</th>
<th>From CK+ to DISFA</th>
<th>From MMI to CK+</th>
<th>From MMI to DISFA</th>
<th>From DISFA to CK+</th>
<th>From DISFA to MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>HTL</td>
<td>Ours</td>
<td>SVM</td>
<td>HTL</td>
<td>Ours</td>
</tr>
<tr>
<td>0.4900</td>
<td>0.2732</td>
<td>0.2234</td>
<td>0.0907</td>
<td>0.5997</td>
<td>0.6572</td>
<td>0.1282</td>
</tr>
<tr>
<td>0.2919</td>
<td>0.3532</td>
<td>0.4114</td>
<td>0.2627</td>
<td>0.3532</td>
<td>0.3936</td>
<td>0.3676</td>
</tr>
<tr>
<td>0.2405</td>
<td>0.1955</td>
<td>0.3453</td>
<td>0.1041</td>
<td>0.1955</td>
<td>0.2536</td>
<td>- -</td>
</tr>
<tr>
<td>0.3453</td>
<td>0.1854</td>
<td>0.3017</td>
<td>0.5342</td>
<td>0.1854</td>
<td>0.3657</td>
<td>- -</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.2669</td>
<td>0.2518</td>
<td>0.3204</td>
<td>0.2106</td>
<td>0.2325</td>
<td>0.2458</td>
</tr>
</tbody>
</table>

TABLE 6

<table>
<thead>
<tr>
<th>AU</th>
<th>From CK+ to UNBC</th>
<th>From MMI to UNBC</th>
<th>From DISFA to UNBC</th>
<th>From UNBC to CK+</th>
<th>From UNBC to MMI</th>
<th>From UNBC to DISFA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>HTL</td>
<td>Ours</td>
<td>SVM</td>
<td>HTL</td>
<td>Ours</td>
</tr>
<tr>
<td>0.4900</td>
<td>0.2732</td>
<td>0.2234</td>
<td>0.0907</td>
<td>0.5997</td>
<td>0.6572</td>
<td>0.1282</td>
</tr>
<tr>
<td>0.2919</td>
<td>0.3532</td>
<td>0.4114</td>
<td>0.2627</td>
<td>0.3532</td>
<td>0.3936</td>
<td>0.3676</td>
</tr>
<tr>
<td>0.2405</td>
<td>0.1955</td>
<td>0.3453</td>
<td>0.1041</td>
<td>0.1955</td>
<td>0.2536</td>
<td>- -</td>
</tr>
<tr>
<td>0.3453</td>
<td>0.1854</td>
<td>0.3017</td>
<td>0.5342</td>
<td>0.1854</td>
<td>0.3657</td>
<td>- -</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.2669</td>
<td>0.2518</td>
<td>0.3204</td>
<td>0.2106</td>
<td>0.2325</td>
<td>0.2458</td>
</tr>
</tbody>
</table>
only model pairwise dependencies, while we use a RBM model to capture global dependencies among AUs.

Compared with SHTL, which only considers a single AU probability given an expression, the proposed method considers both expression-dependent and expression-independent AU relations, which result in better performance on AU activation recognition.

We compare our method with SHTL for cross-database experiments, and the results are listed in Fig. 6. As in the weakly supervised scenarios, both the performances of the proposed method and SHTL maintain a downward trend as the missing rate increases. Our method achieves better performances than SHTL in the experiments of training on the CK+, MMI and DISFA database, particularly when testing...
on the UNBC database, implying that our method is more generalizable than SHTL. Since the emotion setting of the UNBC database is different from the other three databases, the expression-dependent relations under six basic emotions may be not effective for a database of pain emotion setting like the UNBC database. Unlike SHTL which only uses expression-dependent AU probability, we also use expression-independent AU probability, leading to better performances when testing on the UNBC database. For three experiments of training in the UNBC database, the performances of the proposed method are worse than SHTL, but the differences are minor. Specifically, when training on the UNBC database and testing on the CK+ database, the proposed method has close results to SHTL when missing rate is 20, 30, 50, 60, and 90 percent. When training on the UNBC database and testing on the MMI database, the maximum difference is 0.0558 and the minimum of 0.0350. When training on the UNBC database and testing on the DISFA database, the maximum difference is 0.0379 and the minimum of 0.0112. In addition to the labels from the UNBC database, the maximum difference is 0.0112. In addition to the labels from the UNBC database, SHTL trains on another large facial image database with the same expressions as the testing set, while we only train on the UNBC database, a pain expression database.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8249</td>
<td>0.6220</td>
<td>0.8686</td>
<td>0.9129</td>
</tr>
<tr>
<td>2</td>
<td>0.8696</td>
<td>0.7620</td>
<td>0.8547</td>
<td>0.8925</td>
</tr>
<tr>
<td>4</td>
<td>0.7916</td>
<td>0.6910</td>
<td>0.7258</td>
<td>0.7008</td>
</tr>
<tr>
<td>5</td>
<td>0.7347</td>
<td>-</td>
<td>0.7204</td>
<td>0.8189</td>
</tr>
<tr>
<td>6</td>
<td>0.7280</td>
<td>0.7960</td>
<td>0.6147</td>
<td>0.1116</td>
</tr>
<tr>
<td>7</td>
<td>0.5752</td>
<td>0.7910</td>
<td>0.5447</td>
<td>0.4091</td>
</tr>
<tr>
<td>9</td>
<td>0.8794</td>
<td>-</td>
<td>0.8591</td>
<td>0.9036</td>
</tr>
<tr>
<td>12</td>
<td>0.8760</td>
<td>0.7720</td>
<td>0.7265</td>
<td>0.8412</td>
</tr>
<tr>
<td>17</td>
<td>0.8676</td>
<td>0.7430</td>
<td>0.8166</td>
<td>0.8560</td>
</tr>
<tr>
<td>23</td>
<td>0.6727</td>
<td>-</td>
<td>0.5664</td>
<td>0.4157</td>
</tr>
<tr>
<td>24</td>
<td>0.5102</td>
<td>-</td>
<td>0.3529</td>
<td>0.6448</td>
</tr>
<tr>
<td>25</td>
<td>0.9181</td>
<td>-</td>
<td>0.9257</td>
<td>0.9529</td>
</tr>
</tbody>
</table>

Avg. 0.7707 0.7396 0.7147 0.7050

This obvious bias between the training and testing dataset results in poorer performance of the proposed method.

5.2.4 Comparison to State-of-the-Art Supervised Methods with Fully AU-Labeled Data

We compare our weakly supervised learning method to supervised methods with fully AU-labeled data. For within-database experiments, on the CK+ database, we compare the proposed method with MC-LVM [21], STM [36], and HRBM [16]; the results of HRBM are from [21]. On the MMI database, we compare the proposed method with SVM-HMM [37] and FFD [38]. On the DISFA database, we compare the proposed method with MV-LVM [21], HRBM [16], \( l_p \)-MTMKL [40], and iCPM [39]; the results of HRBM and \( l_p \)-MTMKL are from [21]. On the UNBC database, we compare the proposed method with MC-LVM [21], HRBM [16], and \( l_p \)-MTMKL [40]; the results of HRBM and \( l_p \)-MTMKL are from [21]. The comparative results on the four databases are illustrated in Tables 7, 8, 9 and 10 respectively.

As shown in these four tables, our method performs worse than other supervised methods in most cases. This is expected, since other supervised methods train with fully AU-labeled data but we train with expression labels only. Nevertheless, our method achieves comparable or even better performances in some cases. Specifically, on the CK+ database, the average F1-measure of the proposed method is 8.93 percent lower than the best method (MC-LVM), but the proposed method has the close results on AU4, AU12 and AU17, and the best results on AU1, AU2, AU5, AU9, AU24, and AU25. On the MMI database, the average F1-measure of common AUs of the proposed method is

<table>
<thead>
<tr>
<th>AU</th>
<th>SVM-HMM [37]</th>
<th>FFD [38]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5850</td>
<td>0.7270</td>
<td>0.6947</td>
</tr>
<tr>
<td>2</td>
<td>0.7300</td>
<td>0.7270</td>
<td>0.6994</td>
</tr>
<tr>
<td>4</td>
<td>0.6150</td>
<td>0.6930</td>
<td>0.6052</td>
</tr>
<tr>
<td>6</td>
<td>0.5610</td>
<td>0.4850</td>
<td>0.6720</td>
</tr>
<tr>
<td>7</td>
<td>0.6930</td>
<td>0.7370</td>
<td>0.2766</td>
</tr>
<tr>
<td>9</td>
<td>0.3900</td>
<td>0.3640</td>
<td>0.4713</td>
</tr>
<tr>
<td>10</td>
<td>0.8870</td>
<td>0.6920</td>
<td>0.3075</td>
</tr>
<tr>
<td>12</td>
<td>0.7900</td>
<td>0.7590</td>
<td>0.2850</td>
</tr>
<tr>
<td>17</td>
<td>0.7730</td>
<td>0.6220</td>
<td>0.5905</td>
</tr>
<tr>
<td>23</td>
<td>-</td>
<td>0.7650</td>
<td>0.5056</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>0.4120</td>
<td>0.2493</td>
</tr>
<tr>
<td>26</td>
<td>0.7760</td>
<td>0.8470</td>
<td>0.7521</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.6712</td>
<td>0.6652</td>
<td>0.5161</td>
</tr>
<tr>
<td>Avg. of com</td>
<td>0.6712</td>
<td>0.6792</td>
<td>0.5413</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5855</td>
<td>0.3967</td>
<td>0.4221</td>
<td>0.2950</td>
</tr>
<tr>
<td>2</td>
<td>0.6299</td>
<td>0.5592</td>
<td>0.4581</td>
<td>0.2480</td>
</tr>
<tr>
<td>4</td>
<td>0.7285</td>
<td>0.6156</td>
<td>0.4718</td>
<td>0.5680</td>
</tr>
<tr>
<td>6</td>
<td>0.5232</td>
<td>0.5401</td>
<td>0.6279</td>
<td>0.4170</td>
</tr>
<tr>
<td>12</td>
<td>0.8474</td>
<td>0.7916</td>
<td>0.7633</td>
<td>0.7190</td>
</tr>
<tr>
<td>17</td>
<td>0.4863</td>
<td>0.3882</td>
<td>0.4140</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.8160</td>
</tr>
<tr>
<td>26</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.5130</td>
</tr>
</tbody>
</table>

Avg. 0.6335 0.5486 0.5267 0.5109 0.4236

Avg. of com 0.6629 0.5806 0.5492 0.4494 0.4442

As shown in these four tables, our method performs worse than other supervised methods in most cases. This is expected, since other supervised methods train with fully AU-labeled data but we train with expression labels only. Nevertheless, our method achieves comparable or even better performances in some cases. Specifically, on the CK+ database, the average F1-measure of the proposed method is 8.93 percent lower than the best method (MC-LVM), but the proposed method has the close results on AU4, AU12 and AU17, and the best results on AU1, AU2, AU5, AU9, AU24, and AU25. On the MMI database, the average F1-measure of common AUs of the proposed method is
20.3 percent lower than the best method (FFD), but the performances on AU1, AU2, and AU12 are close to other methods and the performances on AU5 and AU7 are better than other methods. On the DISFA database, the average F1-measure of common AUs of the proposed method is 32.99 percent lower than the best method (MC-LVM), but the proposed method has the close results on AU6, AU12, AU17, and AU25. On the UNBC database, the average F1-measure of the proposed method is far below that of the best method (MC-LVM), this is because domain knowledge about the pain expression is deficient, but we have close results on AU4 and AU9. These results indicate that although we train AU classifiers without any AU labels, we can achieve comparable or even better results than supervised methods with fully AU-labeled data in some cases. This demonstrates the effectiveness of leveraging domain knowledge to train AU classifiers. Compared with fully supervised method, the proposed method may have wider applications, since it does not need AU labels.

For cross-database experiments, we compare the proposed method with SVM as shown in Tables 5 and 6. Despite the fact that SVM is a fully supervised method, our method outperforms SVM in all scenarios, demonstrating the better generalization ability of the proposed method. Since there may be biases between databases, fully supervised learning from ground truth labels limits the generalization ability of SVM. While the proposed method uses database independent domain knowledge, and thus has better generalization ability.

6 Conclusion

We propose a new method to learn AU classifiers with only expression labels and without AU labels. Specifically, we summarize domain knowledge about expressions and AUs, and represent it as prior probabilities, including single AU probability given expression as well as expression-independent or expression-dependent conditional probability of one AU under another AU. We sample pseudo AU data for each expression based on the summarized probabilities. We obtain joint prior probability of all considered AUs by training a RBM model and whose input are pseudo AUs sampled from domain knowledge. The AU classifiers are trained by maximizing the log likelihood of AU classifiers with regard to the learned AU label prior. Furthermore, the weakly supervised method can be extended to semi-supervised scenarios with partially labeled data. The within-database experiments demonstrate the effectiveness of our method in learning with domain knowledge. The cross-database experiments further demonstrate the generalization ability of our method.

In this paper, we successfully leverage the summarized domain knowledge between expressions and AUs to train AU activation classifiers from images with expression labels only. The proposed expression-assisted AU analysis method through exploring domain knowledge is not limited to AU activation recognition. Compared with AU activation, AU intensities provide more fine-grained level for facial analysis. Therefore, we will extend the proposed method to AU intensity estimation as long as we have the domain knowledge about expressions and AU intensities in the future.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 61473270, Grant 91748129, Grant 61175037, and Grant 61228304, and in part by the Project from Anhui Science and Technology Agency under Grant 1508085MF223.

References


Authorized licensed use limited to: Rensselaer Polytechnic Institute. Downloaded on January 03,2021 at 04:15:01 UTC from IEEE Xplore. Restrictions apply.


Qiang Ji received the PhD degree in electrical engineering from the University of Washington. He is currently a professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI). From 2010 to 2015, he served as a program director with the National Science Foundation (NSF), Arlington, Virginia, 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, Urbana, Illinois; the Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania; the Department of Computer Science, University of Nevada, Reno, Nevada; and the Air Force Research Laboratory, Rome, New York. He is currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI. His research interests include computer vision, probabilistic graphical models, machine learning, and their applications in various fields. He has published more than 230 papers in peer-reviewed journals and conferences, and has received multiple awards for his work. 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 for numerous international conferences/workshops. He is a fellow of the IEEE and the IAPR.

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 with Rensselaer Polytechnic Institute in Troy, New York. She is currently an associate professor of the 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.

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 interesting is affective computing.

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