Weakly Supervised Facial Action Unit Recognition through Adversarial Training

Guozhu Peng, Shangfei Wang*
University of Science and Technology of China, Hefei, Anhui, China
gzpeng@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract

Current works on facial action unit (AU) recognition typically require fully AU-annotated facial images for supervised AU classifier training. AU annotation is a time-consuming, expensive, and error-prone process. While AUs are hard to annotate, facial expression is relatively easy to label. Furthermore, there exist strong probabilistic dependencies between expressions and AUs as well as dependencies among AUs. Such dependencies are referred to as domain knowledge. In this paper, we propose a novel AU recognition method that learns AU classifiers from domain knowledge and expression-annotated facial images through adversarial training. 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 a weakly supervised AU recognition method via an adversarial process, in which we simultaneously train two models: a recognition model R, which learns AU classifiers, and a discrimination model D, which estimates the probability that AU labels generated from domain knowledge rather than the recognized AU labels from R. The training procedure for R maximizes the probability of D making a mistake. By leveraging the adversarial mechanism, the distribution of recognized AUs is closed to AU prior distribution from domain knowledge. Furthermore, the proposed weakly supervised AU recognition can be extended to semi-supervised learning scenarios with partially AU-annotated images. Experimental results on three benchmark databases demonstrate that the proposed method successfully leverages the summarized domain knowledge to weakly supervised AU classifier learning through an adversarial process, and thus achieves state-of-the-art performance.

1. Introduction

Facial behavior is one of the most important channels for emotional communication between humans. Automatic expression analysis has attracted increasing attention in recent years due to its wide potential in human-robot interaction. Both expression categories and facial action units are adopted to describe facial behavior. For expression categories, Ekman's six universal expression categories (i.e., anger, disgust, fear, happiness, sadness, and surprise) are frequently used. In addition to the six basic expressions, people express many other complex expressions, such as hatred and awe. Some expressions even can not be described verbally. Until now, there have been no established complete expression categories. Unlike expression categories, which only describe several limited facial behaviors, the Facial Action Coding System (FACS) [6] describes facial behavior as combinations of facial action units (AUs), which are related to the contraction of a set of facial muscles. Using FACS, nearly any facial behavior can be deconstructed into the specific AUs and their temporal segments.

Most AU recognition includes supervised learning, in which the fully AU-annotated training images are required.Due to subtle facial appearance changes and significant subject-dependent variations, automatic AU recognition is rather challenging. Many fully AU-annotated facial images would be helpful for automatic AU recognition, but collecting this data is time-consuming and error prone, if not impossible. Since AUs are local, subtle, and difficult to recognize, ground truth AU labels should be provided by qualified FACS experts. Compared to AUs, expression categories are easier to annotate, because expression categories describe global facial behavior, and people recognize global changes more quickly and accurately than local variations. Furthermore, expression categories and facial action units are closely related, since both are used to describe facial behavior. For instance, Du *et al*. [4] found that 99% of the time, people show happiness by raising their cheeks and stretching their mouths. The Emotional Facial Action Coding System (EMFACS) [7] lists emotion-related AU combinations. Prkachin *et al*. [19] found that pain intensity can be inferred from the combination of several AUs (i.e., AU4, AU6, AU7, AU9, AU10 and AU43). In addition to expression-dependent AU relations, there are expression-independent AU relations, such as the co-existence between AU1 and AU2, due to the constraints of facial muscles.

*This is the corresponding author.

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00233

Recently, several researchers have leveraged AU relationships to facilitate AU classifier learning through either generative strategy [26, 23, 14] or discriminative strategy [13, 1, 30, 5, 28]. However, all of these works required fully AU-annotated facial images. A few recent studies focus on AU recognition from partially AU annotated samples. For example, Song *et al*. [22] proposed to encode sparsity and co-occurrent structure of facial action units via compressed sensing and group-wise sparsity inducing priors through a novel Bayesian graphical model. Their method can handle partially observed labels by marginalizing over the unobserved values as a part of the inference procedure. Wu *et al*. [27] and Li *et al*. [15] proposed a multi-label learning method (MLML) that explicitly handles missing labels by enforcing consistency between the predicted labels and the provided labels as well as the local smoothness among the label assignments. Wang *et al*. [25] proposed a Bayesian network model 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. Although AU labels can be partially missing in these works, they still require AU labels to learn AU classifiers.

In this paper, we propose a novel weakly supervised AU recognition method from expression-annotated facial images without any AU labels through adversarial training. Specifically, we simultaneously train two models: a recognition model R that learns AU classifiers, and a discriminative model D that estimates the probability that AU labels generated from domain knowledge rather than the recognized AU labels from R. The training procedure for R is to maximize the probability of D making a mistake. By leveraging an adversarial mechanism, the distribution of recognized AUs is closed to AU prior distribution from domain knowledge. Furthermore, we extend the proposed weakly supervised AU recognition method to semi-supervised learning scenarios when partially AU-annotated images are available. We conduct weakly supervised and semi-supervised experiments on three benchmark databases. The proposed method performs best in most scenarios, demonstrating superiority over state-of-the-art works.

To the best of our knowledge, there is only one related work that can handle AU recognition without AU annotation but with expression labels. Ruiz *et al*. [21] proposed Hidden-Task Learning (HTL) to learn both AU classifiers from images and expression classifiers from AUs without any AU annotations but with extra large-scale facial images labeled with expressions by exploiting domain knowledge about the relation between expressions and AUs. They also extended HTL to Semi-Hidden Task Learning (SHTL) when partial AU annotated samples are provided. Unlike Ruiz *et al*.'s work, which requires both facial images without any annotations and extra large-scale facial images labeled with expressions, our method learns AU classifiers from facial images with expression labels directly and does not need large-scale expression-annotated facial images. Furthermore, since Ruiz *et al*.'s work learns both AU classifiers and expression classifiers, the error caused by expression classifiers may propagate to the AU classifiers. Therefore, we prefer to learn AU classifiers directly. Rather than exploiting expression-dependent domain knowledge only as Ruiz *et al*. did, we exploit both expression-dependent and expression-independent domain knowledge to weakly supervise the learning process of AU classifiers from expression-annotated facial images via an adversarial process.

2. Proposed Method

We propose a weakly supervised AU recognition method from expression-annotated facial images and domain knowledge through adversarial training. First, we summarize a large amount of domain knowledge about AU relationships and sample pseudo AU data based on the summarized domain knowledge. After that, we propose a novel adversarial network for AU recognition, with the goal of making the distribution of AU classifiers' output converge to the distribution of the pseudo AU data generated from domain knowledge. Specifically, the proposed AU recognition adversary network consists of two models: a recognition model R, which learns AU classifiers, and a discriminative model D, which estimates the probability that AU labels generated from domain knowledge rather than the recognized AU labels from R. These two models are trained simultaneously through an adversarial process. The training procedure for R is to maximize the probability of D making a mistake, while the training procedure for D clearly distinguishes the pseudo AU data generated with domain knowledge from the predicted AU labels of the recognition model. By leveraging this adversarial mechanism, the distribution of recognized AUs is closed to AU prior distribution from domain knowledge after training. Furthermore, we extend the proposed weakly supervised AU recognition to semi-supervised learning scenarios when partially AU-annotated images are available by adding a cross-entropy term for the AU-annotated images.

2.1. Summary of Domain Knowledge

To generate pseudo AU data through sampling, we need the expression-dependent and expression-independent probability of AUs. The expression-independent probability is the joint probability of two AUs. Expression-dependent probabilities can be subdivided into two kinds of AU probabilities: the marginal probability of a single AU and the joint probability of multiple AUs. If we generate pseudo AU data based only on the marginal probability of each single AU given an expression, it assumes that all AUs

Table 1: Probabilities of AUs observed in expressions [4].

Expression	Prototypical(and variant AUs)
Anger	4, 7, 24 [10(26%), 17(52%), 23(29%)]
Disgust	9, 10, 17 [4(31%), 24(26%)]
Fear	1, 4, 20, 25 [2(57%), 5(63%), 26(33%)]
Happiness	12, 25 [6(51%)]
Sadness	4, 15 [1(60%), 6(50%), 11(26%), 17(67%)]
Surprise	1, 2, 25, 26 [5(66%)]
pain	4(>50%),6(>50%),7(>50%),9(>50%),10(>50%),43(>50%)

Table 2: Expression-related AU combinations from EMFACS [7, 10].

Expression	AU combinations
Anger	4+5, 4+7, 4+5+7, 17+24,23
Fear	1+2+4, 20
Disgust	9, 10(only)
Happiness	12, 6+12, 7+12
Sadness	1or1+4, 15, 6+15, 11+17, 11+15
Surprise	1+2+5(low), 1+2+26, 1+2+5(low)+26

Table 3: Expression-independent AU relations [6, 14].

coexistent	mutually exclusive
AU1—AU2—AU5	AU12—AU15, AU12—AU17
AU4—AU7—AU9	AU2—AU6, AU2—AU7, AU2—AU9
AU15—AU17—AU24	AU15—AU25, AU17—AU25
AU23—AU24	AU23—AU25, AU24—AU25

are independent to each other given that expression. This is unreasonable. Therefore, we need both types of expression-dependent probabilities.

For expression-dependent AU relations, we first consider the domain knowledge about the six basic expressions and AUs. For marginal probability of single AU given expressions, we adopt the observations from [4], as shown in Table 1. In Table 1, the percentage in parentheses following the AU is the marginal probability of a single AU given the expressions. The expression-dependent marginal probability of AUs which are not followed by parentheses is larger than 70%. The expression-dependent marginal probability of AUs which are not listed is less than 20%. For example, the fourth row of Table 1 means: P(AU12 | happiness)$\geq$ 70%, P(AU25 | happiness)$\geq$ 70%, P(AU6 | happiness)$=$ 51% and P(AU4 | happiness)$<$ 20%. These probabilities can provide weak supervisory information for classifiers' training.

For joint probability of multiple AUs given expressions, we adopt the domain knowledge from the Emotion Facial Action Coding System (EMFACS) [7] as shown in Table 2. Table 2 lists AU combinations for each expression. These combinations are co-existent relations among AUs. For example, though AU 6 and AU12 don't correlate to each other according to facial anatomy, they almost always appear simultaneously during happiness,i.e., P(AU6 |AU12, happiness) is very large.

In addition to the domain knowledge about six basic expressions, we also have domain knowledge about pain expression. In 1992, Prkachin *et al*. [19] found 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. Prkachin *et al*. [20] explicitly defined the Prkachin and Solomon pain intensity(PSPI) as shown in Eq. 1:

$$\mathrm{PSPI} = \mathrm{AU4} + (\mathrm{AU6orAU7}) + (\mathrm{AU9orAU10}) + \mathrm{AU43} \quad (1)$$

Eq. 1 suggests that these six AUs play a significant role in pain expression. So for pain frames, we think the occurrence probability of each of these AUs should be higher than 50%, as shown in Table 1, For example, P(AU4|pain)>50%.

Expression-independent AU relations include both co-existent and mutually exclusive relations. These expression-independent AU probabilities are mainly caused by muscular structure of human face. For example, *inner brow raiser* (AU1) and *outer brow raiser* (AU2) are both related to the muscle group *Frontalis*. Most people cannot make a facial movement of AU1 without AU2, and vice versa. It means both P(AU2 |AU1) and P(AU1 |AU2) are very large. This is a kind of co-existent relation. *Lip Corner Puller* (AU12) rarely appears with *Lip Corner Depressor* (AU15). The former is produced by the muscle group *zygomaticus major*, and the latter is produced by the muscle group *depressor anguli oris*. It means P(AU15 |AU12) and P(AU12 |AU15) are very small. This is referred as a mutually exclusive relation. These relations are caused by the mechanism of muscles and are universal for all expressions. Therefore, they are referred as expression-independent AU probabilities. Table 3 lists expression-independent AU relations summarized from [6, 14]. Take the first row of Table 3 for example, AU1 and AU2 are co-existent, AU12 and AU25 are mutually exclusive.

2.2. Pseudo AU Label Generation

After summarizing domain knowledge including expression-dependent and expression-independent AU probabilities, we must translate the domain knowledge into pseudo AU labels. We generate pseudo AU labels for each expression, since expression annotations are available during training. Table 1 provides expression-dependent AU probabilities, therefore we sample first AU with P(AU1| Expression) from Table 1. For the AUs with concrete expression dependent AU probabilities, such as P(AU17 | anger)=0.52, we adopt their probabilities listed in Table 1 directly. For AUs whose expression dependent probabilities larger than 0.70, such as P(AU4 |anger)$\geq$ 70%, we sample the probability parameter in $[0.7, 1]$. For AUs whose expression dependent probabilities less than 0.20, such as P(AU1 | anger)$<$ 20%, we sample the probability parameter in $[0, 0.2]$. The process for the rest of the AUs depends on whether the AU has relation with existing

AUs. If not, we still generate pseudo AU labels according to its expression-dependent marginal probability listed in Table 1. If there is a relation, we generate pseudo AU labels according to the coexistence relations and mutual exclusive relations with existed AUs, as listed in Table 2 and Table 3 respectively. For expression-dependent and expression-independent AU probabilities listed in Table 2 and Table 3, we sample the probability parameter in $[0.7, 1]$ with co-existence and $[0, 0.2]$ with mutual exclusion, following the same criterion of Table 1. The detailed sampling algorithm is shown in Algorithm 1.

2.3. AU Recognition Adversary Network (RAN)

We propose a novel adversary network for AU recognition inspired by the generative adversarial network (GAN) [9]. Given a training set $\mathbf{X} = \{(X_i, \mathrm{E}_i)\}_{i=1}^m$, where $X_i \in \mathbb{R}^d$ represents the sample features and $\mathrm{E}_i \in \{1, 2, ..., P\}$ is the expression label. Our goal is to train AU classifiers from expression labels and domain knowledge without any AU labels. Therefore, the object is to make sure that the output of AU classifiers are consistent with our domain knowledge. The distribution of output P_{Y_c} should converge to the distribution of pseudo AU data P_{Y_s}.

The framework of the proposed AU recognition adversary network is shown in Figure 1. Instead of including a generator G like GAN, the proposed RAN includes a recognition model R, as shown in Part 1 of Figure 1. The recognition model R is an AU classifier to recognize AUs from facial images. The recognized AU labels are the adversarial samples and are regarded as "fake". As shown in Part 2 of Figure 1, the pseudo AU data generated through domain knowledge are regarded as "real". The input of the discriminator D includes both the expression label and the AU labels (either pseudo AU labels or recognized AU labels). The expression label is used as a switch to select the corresponding D_i. Without loss of generality, assume we consider P expression categories in this paper. Therefore, as shown in Part 3 of Figure 1, P discriminators, $\{D_1, D_2, ..., D_P\}$ are used to distinguish "fake" from "real" AU labels for P expressions respectively, since the AU distributions are dependent on expression labels. The objective is given as:

$$\min_R \max_{D_1, D_2, ..., D_P} \sum_{\mathrm{E}=1}^{P} [\mathbb{E}_{Y \sim P_{Y_s}(Y|\mathrm{E})}[\log D_{\mathrm{E}}(Y)] + \mathbb{E}_{X \sim P_X(X|\mathrm{E})}[\log(1 - D_{\mathrm{E}}(R(X)))]], \tag{2}$$

where $P_X(X|E)$ represents the distribution of facial images with expression label E. Since we train D and R alternately, we define objective $\mathcal{L}_D$ for P discriminators and $\mathcal{L}_R$ for AU classifiers as Eqs. 3 and 4. In practice, it's better for R to maximize $\log(D(R(X)))$ instead of minimizing $\log(1 - D(R(X)))$ [8]. The training process is described as Algorithm 2.

$$\mathcal{L}_D = \min_{D_1, D_2, ..., D_P} -\sum_{\mathrm{E}=1}^{P} [\mathbb{E}_{Y \sim P_{Y_s}(Y|\mathrm{E})}[\log D_{\mathrm{E}}(Y)] + \mathbb{E}_{X \sim P_X(X|\mathrm{E})}[\log(1 - D_{\mathrm{E}}(R(X)))]], \tag{3}$$

$$\mathcal{L}_R = \min_R -\sum_{\mathrm{E}=1}^{P} \mathbb{E}_{X \sim P_X(X|\mathrm{E})}[\log D_{\mathrm{E}}(R(X))]. \tag{4}$$

Any classifier can be used in our RAN model. For the discriminator, we use a three-layer feedforward net, and ReLu for the hidden layer and sigmoid for the output layer. For the recognition model, we use a linear function as follows:

$$R(X) = f_\theta(X) = \sigma(WX + b), \tag{5}$$

where σ is the sigmoid function and $\theta = \{W, b\}$ are parameters to be trained. For the optimization method, any gradient-based learning rule could be used to update parameters. We use the ADAM [11] algorithm in this paper.

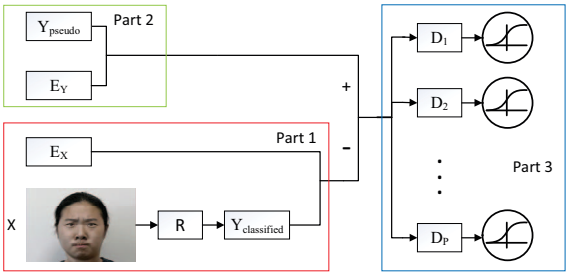

Figure 1: The framework of RAN. In Part 1, the facial feature X is inputted into recognizer R and get the "fake" AU vector, the "real" AU data generated in section 2.2 are in Part 2. In part 3, P discriminators are trained, "real" or "fake" AU data are inputted to corresponding discriminator with the same expression label. See text for details.

Algorithm 1 The sampling of pseudo AU data.

Input: The domain knowledge about expressions and AUs listed in Tables 1, 2 and 3, and sampling size N.
Output: The pseudo AU samples.

```
for expression E_p (p = 1, 2, ..., P) do
  repeat
    generate sample Y_i = {y_i^1, y_i^2, ..., y_i^L}.
    sample first AU with P(AU_1|E) from Table 1.
    for j-th(j = 2, 3, ..., L) component y_i^j do
      if any AU_s related to AU_j is already generated
      and this relation is from Table 2 then
        sample AU_j with P(AU_j|AU_s, E).
      else if this relation is from Table 3 then
        sample AU_j with P(AU_j|AU_s).
      else
        sample AU_j with a P(AU_j|E) from Table 1.
      end if
    end for
  until we have already got N samples.
end for
```

2.3.1 Extension to Semi-Supervised Learning

We extend the proposed weakly supervised AU recognition method to semi-supervised learning when partial AU anno-

tations are available. Let $\mathbf{X^S} = \{(X_i, Y_i)\}_{i=1}^{n} (n \leq m)$ is a subset of $\mathbf{X}$ that is annotated with AU labels, where $Y_i \in \{0,1\}^L$ and L is the number of AUs. For examples with AU labels, a cross-entropy term is incorporated into the objectives from Eq. 4. The cross-entropy loss for data pair (X_i, Y_i) is given as:

$$\mathcal{L}_{CE}(X_i, Y_i) = -\left[Y_i^T \log R(X_i) + (\mathbf{1} - Y_i)^T \log[\mathbf{1} - R(X_i)]\right]. \quad (6)$$

Then, the updated objective $\mathcal{L}_R^S$ for classifier R in semi-supervised RAN are given as Eq. 7, where $\alpha \in [0,1]$ is a trade-off between two terms. The objective of discriminator D has not changed: $\mathcal{L}_D^S = \mathcal{L}_D$.

$$\begin{aligned}\mathcal{L}_R^S = \min_R (1-\alpha)\mathbb{E}_{(X,Y)\sim\mathbf{X^S}}\mathcal{L}_{CE}(X,Y) - \\ \alpha \sum_{\mathrm{E}=1}^{P} \mathbb{E}_{X\sim P_X(X|\mathrm{E})}[\log D_{\mathrm{E}}(R(X))]\end{aligned} \quad (7)$$

3. Experiments

3.1. Experimental Conditions

In our experiments, one posed and two spontaneous databases are used: the Extended Cohn-Kanade database (CK+) [16], the MMI database [18], and the UNBC-McMaster Shoulder Pain Expression Archive database [17].

The CK+ database contains 593 sequences from 123 subjects performing posed expressions. Among them, we use 309 sequences of 106 subjects that are annotated with six basic expressions and AUs; 12 AUs (1, 2, 4, 5, 6, 7, 9, 12, 17, 23, 24, 25), whose frequency of occurrence is larger than 10 % are considered. The MMI database consists of 2900 videos and images from 75 subjects. As with the CK+ database, the sequences annotated with six basic expressions and AUs are chosen, and we obtain 171 sequences from 27 subjects. 13 AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 17, 23, 25, 26) with occurrence frequency larger than 10% are considered in our work. The UNBC database consists of 200 video recordings of 25 different patients suffering from shoulder pain. Each frame is code with PSPI. In this paper, we define the frames with PSPI>4 as "pain" and frames with PSPI=0 as "no pain". From 30 sequences of 17 subjects where exist pain frames, we select all pain and no pain frames, 7319 frames in total. Six AUs (4, 6, 7, 9, 10, 43) associated with pain are considered.

Algorithm 2 Adversarial weakly supervised AU learning.

Input: The training images with expression labels, pseudo AU data, max number of training step K, update number of D per step (T_D), update number of R per step (T_R).

Output: The AU recognizer R.

Initialize parameters of classifier and six discriminators.
for $k = 1, 2, ..., K$ **do**
  for $t = 1, ..., T_D$ **do**
    Sample minibatch of m samples (x^1, E_{x^1}), $(x^2, \mathrm{E}_{x^2}), ..., (x^m, \mathrm{E}_{x^m})$ from training images.
    Sample minibatch of m samples (y^1, E_{y^1}), $(y^2, \mathrm{E}_{y^2}), ..., (y^m, \mathrm{E}_{y^m})$ from pseudo AU data.
    Update six discriminators by descending its gradient:
    $$\nabla_{\theta_d} - \frac{1}{m}\sum_{i=1}^{m}[\log D_{\mathrm{E}_{y^i}}(y^i) + \log(1 - D_{\mathrm{E}_{x^i}}(R(x^i)))]$$
  end for
  for $t = 1, ..., T_R$ **do**
    Sample minibatch of m samples (x^1, E_{x^1}), $(x^2, \mathrm{E}_{x^2}), ..., (x^m, \mathrm{E}_{x^m})$ from training images.
    Update the classifier by descending its gradient:
    $$\nabla_{\theta_r} - \frac{1}{m}\sum_{i=1}^{m}\log D_{\mathrm{E}_{x^i}}(R(x^i))$$
  end for
end for

For weakly 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 works, i.e., HTL [21], the only work which recognizes AUs without AU-labeled images. For the MMI database, we conduct AU recognition with the implementation of the HTL method, since Ruiz *et al*. [21] does not provide experimental results on the MMI database. For other databases, we directly compare our results to the experimental results listed in [21]. For semi-supervised learning scenarios, we randomly miss AU labels with certain probabilities, i.e., 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, and conduct AU recognition experiments for each missing rate five times. We compare our work with state-of-the-art works, including BGC-S [22], MLML [27], BN [25] and SHTL [21]. These works can recognize AUs from partially AU-labeled images. Since the experimental conditions of these works are different from ours, we re-conduct the experiments using the provided code. Specifically, Song *et al*. [22] only provided the result under missing 50% labels, not nine missing rates; Wu *et al*. [27] used four missing rates, (i.e., 20%, 40%, 60% and 80%) and adopted different performance metrics; Wang *et al*. [25] conducted semi-supervised experiments by missing one certain AU with a specific proportion (unlike our experiment, which misses all AUs of one image), and Ruiz *et al*. [21] didn't provide the results of nine missing rates ei-

Facial feature points are used for feature. On the CK+ database and the UNBC database, the feature points are provided by the database constructors. On the MMI database, the feature points are extracted with IntraFace [3]. All feature points are normalized, so that the eye centers fall on the given positions for all images based on affine transformation. We report F1 score as the performance measure.

We conduct experiments of both AU recognition from facial images with expression labels only (i.e., weakly supervised learning) and AU recognition from 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 works, i.e., HTL [21], the only work which recognizes AUs without AU-labeled images. For the MMI database, we conduct AU recognition with the implementation of the HTL method, since Ruiz *et al*. [21] does not provide experimental results on the MMI database. For other databases, we directly compare our results to the experimental results listed in [21]. For semi-supervised learning scenarios, we randomly miss AU labels with certain probabilities, i.e., 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, and conduct AU recognition experiments for each missing rate five times. We compare our work with state-of-the-art works, including BGC-S [22], MLML [27], BN [25] and SHTL [21]. These works can recognize AUs from partially AU-labeled images. Since the experimental conditions of these works are different from ours, we re-conduct the experiments using the provided code. Specifically, Song *et al*. [22] only provided the result under missing 50% labels, not nine missing rates; Wu *et al*. [27] used four missing rates, (i.e., 20%, 40%, 60% and 80%) and adopted different performance metrics; Wang *et al*. [25] conducted semi-supervised experiments by missing one certain AU with a specific proportion (unlike our experiment, which misses all AUs of one image), and Ruiz *et al*. [21] didn't provide the results of nine missing rates ei-

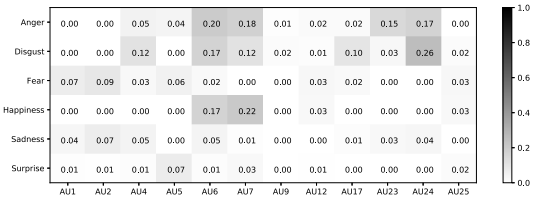

Figure 2: Evaluation of Adversarial Learning. The absolute difference between the distribution of pseudo AU labels and the distribution of recognized AU labels from R model.

ther. Furthermore, none of them conducted experiments on the MMI database.

3.2. Evaluation of Adversarial Learning

To evaluate the adversarial learning process, we compare the distribution of pseudo AU labels generated from domain knowledge and the distribution of recognized AU labels from R model. Take experiments on the CK+ database for example, Figure 2 illustrates the absolute difference between the distribution of 12 pseudo AUs generated from domain knowledge and the distribution of recognized AU labels from R model for the training set when the proposed RAN reaches convergence. From Figure 2, we can find that the differences are very small, ranging from 0.04 to 0.26. This demonstrates that the adversarial learning process successfully makes the distribution of recognized AU labels close to the distribution of pseudo AU labels. Therefore, the proposed RAN effectively leverages domain knowledge to provide weak supervision for AU recognition from expression-annotated facial images. It leads to superior performance of the proposed method to state of the art work in the following sections.

3.3. Weakly Supervised AU Recognition

The within-database experimental results are illustrated in Table 4. From Table 4, we find that RAN outperforms HTL on all three databases, with a higher average F1 score and higher F1 scores for most AUs. Specifically, on the CK+ databases, the average F1 score of common AUs of RAN is 0.7691, achieving a 22.8% improvement over HTL. For specific AUs, the F1 scores of RAN are higher than HTL on 7 out of 10 AUs. On AU1, AU2 and AU9, this improvement is more than 50%. On the MMI database, the average F1 score of RAN is 0.5206, achieving a 20.8% improvement over HTL. For specific AUs, RAN achieves better performance on 10 out of 13 AUs; on AU1, AU4, and AU17, the improvement is more than 50%. On the UNBC database, the average F1 score of common AUs of RAN is 0.3365, achieving more than double the performance of HTL. For all specific AUs except AU 10, the F1 score of RAN are higher than HTL. The results above strongly demonstrate that the proposed method not only works well on posed facial expressions (in the CK+ database), but also works well on spontaneous facial expression. Especially the better performance on the UNBC database, which is

Table 4: Within-database experiment results (F1) of weakly supervised AU recognition.

AU	CK+		MMI		UNBC	
	HTL	RAN	HTL	RAN	HTL	RAN
1	0.6190	**0.9365**	0.3857	**0.6754**	-	-
2	0.4510	**0.8987**	**0.6000**	0.5967	-	-
4	**0.8160**	0.7414	0.3407	**0.6097**	0.0830	**0.4416**
5	0.7480	**0.7989**	0.5882	**0.7092**	-	-
6	**0.5870**	0.5337	**0.3922**	0.3430	0.2950	**0.5019**
7	0.3270	**0.4465**	0.3220	**0.4258**	0.1720	**0.3632**
9	0.4870	**0.8861**	0.3495	**0.4088**	0.0530	**0.3068**
10	-	-	**0.4118**	0.2355	**0.0850**	0.0689
12	**0.8520**	0.8292	0.5439	**0.6881**	0.4260	-
17	0.6750	**0.6789**	0.2167	**0.5110**	-	-
23	-	0.4122	0.2178	**0.2239**	-	-
24	-	0.4879	-	-	-	-
25	0.7010	**0.9369**	0.6631	**0.7025**	0.1240	-
26	-	-	0.5714	**0.6379**	0.1500	-
43	-	-	-	-	-	0.5754
Avg.	0.6263	**0.7156**	0.4310	**0.5206**	0.1735	**0.3763**
Avg. of com	0.6263	**0.7691**	0.4310	**0.5206**	0.1367	**0.3365**

a database of non-basic emotion setting, suggests that our method will be not limited to basic emotion settings. As long as there exists domain knowledge about the considered expression, the proposed method can achieve good results.

Unlike HTL, which uses expression-dependent AU relations, the proposed RAN employs both expression-dependent and expression-independent AU relations, like the expression-independent joint probability of two AUs (e.g., AU1/AU5 and AU7/AU9). These expression-independent relations provide more structure information of AUs, and result in better performance. Furthermore, HTL trains one classifier from AU to expression and another classifier from feature to AU separately. Any error caused by expression classifiers may propagate to the AU classifiers. While the proposed RAN learns classifiers and discriminators simultaneously through adversarial process, thus avoiding error propagation. As discussed in Section 2.3, the adversarial learning process successfully approximates the distribution of recognized AU labels so that it is close to the distribution of pseudo AUs generated from domain knowledge, and effectively leverages domain knowledge to provide weak supervision for AU recognition from expression-annotated facial images. This leads to better performance.

We compare our method to HTL for cross-database experiments. The results are listed in Table 5. Our method outperforms HTL in first four experiments, i.e., training on the CK+ database and testing on the MMI database, training on the CK+ database and testing on the UNBC database, training on the MMI database and testing on the CK+ database, and training on the MMI database and testing on the UNBC database. Our method performs especially well on experiments that test on the UNBC database. These results demonstrate the proposed method successfully leverages more complete domain knowledge for better performance. For experiments that train on the UNBC database

Table 5: Cross-database experiment results (F1) of weakly supervised AU recognition.

AU	From CK+ to MMI			From CK+ to UNBC			From MMI to CK+			From MMI to UNBC			From UNBC to CK+			From UNBC to MMI		
	SVM	HTL	RAN	SVM	HTL	RAN	SVM	HTL	RAN	SVM	HTL	RAN	SVM	HTL	RAN	SVM	HTL	RAN
1	**0.6463**	0.3857	0.5455	-	-	-	0.6215	0.6190	**0.7459**	-	-	-	-	-	-	-	-	-
2	0.5546	0.6000	**0.6748**	-	-	-	0.6638	0.4510	**0.7845**	-	-	-	-	-	-	-	-	-
4	0.3857	0.3407	**0.6395**	0.1900	0.0830	**0.2572**	0.5799	**0.8160**	0.5959	0.0987	0.0830	**0.1647**	0.2412	**0.8160**	0.4470	0.2804	**0.3407**	0.2927
5	0.4355	0.5882	**0.6316**	-	-	-	0.7393	**0.7480**	0.6279	-	-	-	-	-	-	-	-	-
6	0.2811	**0.3922**	0.3485	0.2919	0.2950	**0.4526**	0.5244	**0.5870**	0.5240	0.2627	**0.2950**	0.2899	0.3438	**0.5870**	0.4066	0.2432	**0.3922**	0.2727
7	0.2373	0.3220	**0.4430**	0.2405	0.1720	**0.4311**	0.4397	0.3270	**0.4504**	0.1041	**0.1720**	0.0123	0.1522	0.3270	**0.4036**	0.2198	**0.3220**	0.2833
9	**0.4127**	0.3495	0.3288	**0.3453**	0.0530	0.0381	0.3907	**0.4870**	0.4298	**0.5342**	0.0530	0.4543	0.2899	**0.4870**	0.1714	**0.4308**	0.3495	0.3800
10	-	-	-	-	-	-	-	-	-	0.0534	0.0850	**0.3212**	-	-	-	0.1111	**0.4118**	0.2376
12	0.4156	**0.5439**	0.5354	-	-	-	0.4434	**0.8520**	0.7634	-	-	-	-	-	-	-	-	-
17	0.1765	0.2167	**0.4478**	-	-	-	0.3636	**0.6750**	0.6543	-	-	-	-	-	-	-	-	-
23	0.0702	0.2178	**0.2203**	-	-	-	**0.2206**	-	0.2037	-	-	-	-	-	-	-	-	-
25	**0.7984**	0.6631	0.7577	-	-	-	0.6870	0.7010	**0.8421**	-	-	-	-	-	-	-	-	-
Avg. of com	0.4014	0.4200	**0.5066**	0.2669	0.1508	**0.2948**	0.5453	0.6263	**0.6418**	0.2106	0.1376	**0.2482**	0.2567	**0.5543**	0.3572	0.2571	**0.3632**	0.2933

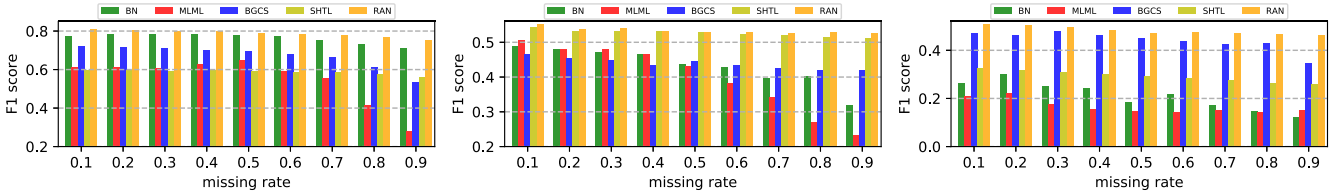

Figure 3: Within-database experiments (F1)of semi-supervised AU recognition. Left: results on the CK+ database, Middle: results on the MMI database, Right: results on the UNBC database.

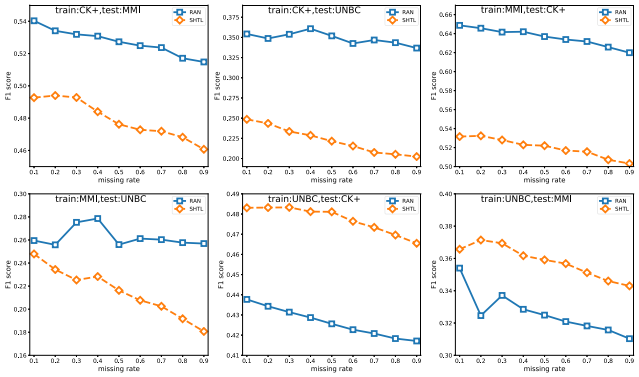

Figure 4: Cross-database experiment results (F1) of semi-supervised AU recognition.

and test on the CK+ or MMI database, HTL achieves better performance than RAN. This is because HTL trains on a large facial image database, which has the same emotion setting as the testing set. While the proposed method only trains on the UNBC database, which is a database of pain emotion setting. Therefore, when testing on the databases of basic emotion setting like the CK+ and MMI databases, the performance of the proposed method is worse than HTL.

3.4. Semi-Supervised AU Recognition

The within-database experimental results of semi-supervised scenarios are listed in Figure 3. From Figure 3, we find that on all three databases, the performance of the proposed method decreases as the missing rate increases. This is expected, since more AU labels provide more accurate information for AU classifiers.

Compared with four state-of-the-art semi-supervised AU recognition methods, we find that RAN performs best in most scenarios, which demonstrates the superiority of the proposed method in handling missing AUs for AU recognition. MLML handles missing labels by leveraging the label consistency and label smoothness, but the smoothness between labels isn't always correct. BGCS naturally handles partially observed labels by marginalizing over the unobserved values as a part of the inference procedure. BN adopts expression labels to assist in AU classifier training and complements the missing AU by capturing the relations between facial expression and AUs. All methods mentioned above handle missing AU labels through AU relations or AU-expression relations learned from partially available ground-truth labels, while the proposed method learns AU relations from pseudo data generated from summarized domain knowledge through adversarial training. The AU relations coded in domain knowledge are more general than those embedded in ground truth AU labels, and lead to better performance. SHTL leverages the expression-dependent probability of a single AU summarized from domain knowledge to handle missing labels. However, SHTL only considers single AU probability given expression, while the proposed method considers both expression-dependent AU relations and expression-independent AU relations. The more complete AU relations employed in the proposed method result in better performance on AU recognition.

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

Table 6: Comparison to state-of-the-art supervised methods with fully AU labeled data on three databases.

Database	Method	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU17	AU23	AU24	AU25	AU26	AU 43	Avg.
CK+	MC-LVM [5]	0.8249	0.8696	**0.7916**	0.7347	0.7280	0.5752	0.8794	-	**0.8760**	**0.8676**	**0.6727**	**0.5102**	0.9181	-	-	**0.7707**
	STM [2]	0.6220	0.7620	0.6910	-	**0.7960**	**0.7910**	-	-	0.7720	0.7430	-	-	-	-	-	0.7396
	HRBM [26]	0.8686	0.8547	0.7258	0.7204	0.6147	0.5447	0.8591	-	0.7265	0.8166	0.5664	0.3529	0.9257	-	-	0.7147
	RAN	**0.9365**	**0.8987**	0.7414	**0.7989**	0.5337	0.4465	**0.8861**	-	0.8292	0.6789	0.4122	0.4879	**0.9369**	-	-	0.7156
MMI	SVM-HMM [24]	0.5850	**0.7300**	0.6150	0.5610	0.6930	0.3900	**0.8870**	**0.7900**	**0.7730**	-	-	-	0.7760	0.5830	-	**0.6712**
	FFD [12]	**0.7270**	0.7270	**0.6930**	0.4850	**0.7370**	0.3640	0.6920	0.7590	0.6220	**0.7650**	**0.4120**	-	**0.8470**	**0.8180**	-	0.6652
	RAN	0.6754	0.5967	0.6097	**0.7092**	0.3430	**0.4258**	0.4088	0.2355	0.6881	0.5110	0.2239	-	0.7025	0.6379	-	0.5206
UNBC	MC-LVM [5]	-	-	**0.4720**	-	**0.9775**	0.6788	**0.3713**	**0.5823**	-	-	-	-	-	-	**0.7251**	**0.6345**
	HRBM [26]	-	-	**0.4720**	-	0.9393	0.6367	0.2980	0.5239	-	-	-	-	-	-	0.6954	0.5942
	l_p-MTMKL [29]	-	-	0.3769	-	**0.9775**	**0.7008**	0.3328	0.4179	-	-	-	-	-	-	0.4403	0.5410
	RAN	-	-	0.4416	-	0.5019	0.3632	0.3068	0.0689	-	-	-	-	-	-	0.5754	0.3763

the CK+ and the MMI database are better than performances on the UNBC database. Both the CK+ database and the MMI database consist of facial images with six basic expressions, while the UNBC database is a database of pain expression. The relations between AUs and six basic expressions have been studied more thoroughly than the relations between AUs and pain expression, so the reviewed domain knowledge of six basic expressions provides more detailed and concrete expression dependent AU probabilities than the reviewed domain knowledge of pain expression. This clearer domain knowledge provides better supervision for AU recognition.

For cross-database experiments, we compare our method with SHTL, and the results are listed in Figure 4. As in the weakly supervised scenarios, our method achieves better performances than SHTL in first four experiments. This implies that our method is more generalizable that SHTL. However, in the last two experiments (which train in the UNBC database), the performances of RAN are worse than SHTL. In addition to the labels from the UNBC database, SHTL trains on another large facial image database that has the same emotion setting as the testing set (i.e. basic emotions setting). While the proposed method only trains on the UNBC database, which is a pain expression database, and has a different emotion setting with testing set.

3.5. Comparison to State-of-the-art Supervised Methods with Fully AU-Labeled Data

Our weakly supervised learning method is also compared with supervised methods with fully AU-labeled data. For within-database experiments, on the CK+ database, we compare RAN to MC-LVM [5], STM [2] and HRBM [26], the results of HRBM are collected from [5]. On the MMI database, we compare RAN with SVM-HMM [24] and FFD [12]. On the UNBC database, we compare RAN with MC-LVM [5], HRBM [26] and l_p-MTMKL [29]. We use the results of HRBM and l_p-MTMKL in [5]. The comparisons on the three database are shown in Table 6.

From Table 6, we find our method performs worse than other supervised methods in most cases. This is reasonable, because we train with expression labels only while other supervised methods train with fully AU-labeled data. Yet even so, the results of our method in some cases are comparable or even better. Specifically, on the CK+ database, the average F1 score of RAN is 0.7156, which is 7.15% lower than the best method (MC-LVM), but is 0.13% higher than HRBM, and RAN has the best results on AU1, AU2, AU5, AU9 and AU25. On the MMI database, the average F1 score of RAN is 22.44% lower than the best method (SVM-HMM), but the performances on AU5 and AU7 are better than other methods. On the UNBC database, the proposed method has close results to compared methods on AU4, AU9 and AU43. These results demonstrate that even though we train AU classifiers without any AU labels, we achieve comparable or even better results than supervised methods with fully AU-labeled data, demonstrating the effectiveness of the proposed method which leverages domain knowledge through an adversarial mechanism.

For cross-database experiments, we compare the proposed RAN with SVM, and the results are listed in Table 5. From this table, we find that our method outperforms SVM in all scenarios, although SVM is a fully supervised method. Fully supervised learning from ground truth labels limits the generalization ability of SVM due to database bias. While the proposed method uses domain knowledge, which is not dependent on databases.

4. Conclusion

We propose a novel weakly supervised AU recognition method to learn AU classifiers with only expression labels. Specifically, we notice that there exist domain knowledge about expressions and AUs that can be represented as prior probabilities. We generate pseudo AU data for each expression; for AU classifiers' training, we propose an RAN model, which consists of a recognition model and a discrimination mode trained simultaneously by leveraging an adversarial process, to make the distribution of the recognized AU close to the distribution of the pseudo AU data. Furthermore, we extend the proposed method to semi-supervised learning with partially AU-annotated images. Both weakly supervised and semi-supervised experiments demonstrate the effectiveness of the proposed method.

Acknowledgements: This work was supported by the National Science Foundation of China (Grant No. 91748129, 61473270, 61175037, 61228304), and the project from Anhui Science and Technology Agency (1508085SMF223).

References

[1] T. Almaev, B. Martinez, and M. Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3774–3782, 2015.

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[3] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. Intraface. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[4] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[5] 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.

[6] E. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

[7] W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 1983.

[8] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] J. C. Hager. A comparison of units for visually measuring facial actions. *Behavior Research Methods, Instruments, & Computers*, 17(4):450–468, 1985.

[11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010.

[13] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] Y. Li, J. Chen, Y. Zhao, and Q. Ji. Data-free prior model for facial action unit recognition. *Affective Computing, IEEE Transactions on*, 4(2):127–141, April 2013.

[15] 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.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[17] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[18] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[19] K. M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, 1992.

[20] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008.

[21] 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.

[22] 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.

[23] Y. Tong and Q. Ji. Learning bayesian networks with qualitative constraints. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[24] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2012.

[25] J. Wang, S. Wang, and Q. Ji. Facial action unit classification with hidden knowledge under incomplete annotation. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 75–82. ACM, 2015.

[26] 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.

[27] 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.

[28] X. Zhang and M. H. 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. IEEE, 2014.

[29] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn. A l p-norm mtmkl framework for simultaneous detection of multiple facial action units. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1104–1111. IEEE, 2014.

[30] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.