Classifier Learning with Prior Probabilities for Facial Action Unit Recognition

Yong Zhang[1,2], Weiming Dong[1], Bao-Gang Hu[1], and Qiang Ji[3]*
[1]National Laboratory of Pattern Recognition, CASIA
[2]University of Chinese Academy of Sciences
[3]Rensselaer Polytechnic Institute
zhangyong201303@gmail.com, weiming.dong@ia.ac.cn, hubg@nlpr.ia.ac.cn, qji@ecse.rpi.edu

Abstract

Facial action units (AUs) play an important role in human emotion understanding. One big challenge for data-driven AU recognition approaches is the lack of enough AU annotations, since AU annotation requires strong domain expertise. To alleviate this issue, we propose a knowledge-driven method for jointly learning multiple AU classifiers without any AU annotation by leveraging prior probabilities on AUs, including expression-independent and expression-dependent AU probabilities. These prior probabilities are drawn from facial anatomy and emotion studies, and are independent of datasets. We incorporate the prior probabilities on AUs as the constraints into the objective function of multiple AU classifiers, and develop an efficient learning algorithm to solve the formulated problem. Experimental results on five benchmark expression databases demonstrate the effectiveness of the proposed method, especially its generalization ability, and the power of the prior probabilities.

1. Introduction

Automatic facial action unit (AU) recognition has attracted increasing attention, and achieved impressive progress in the past decades [2, 7, 33, 3]. Due to subtle facial appearance changes, significant across-subject variations, uncertain and ambiguous facial motion measurements, and lack of ground-truth AU annotations, AU recognition is very challenging. Recently, several works turn to leverage relationships among AUs or relationships between expression and AUs to facilitate AU classifier learning, since there exist significant dependencies among multiple AUs. They either use discriminative approaches or generative approaches. For discriminative approaches, the relationships are leveraged to introduce regularization on the parameters of classifiers [32, 12]. They encourage the parameters of AU classifiers or the estimated AU labels to be close if the AUs are positively correlated. For generative approaches, the relationships are directly learned from training AU labels and then applied to predict AUs [20, 25]. Although the dependencies among AUs can improve the performance of AU classifiers, almost all of these works formulate AU recognition as a supervised learning or semi-supervised learning, and thus require fully AU-annotated or partly-annotated facial images for training. However, the AU annotation needs strong domain expertise.

Can we learn AU classifiers without AU annotations? In this paper, we propose a novel method to jointly learn classifiers for multiple AUs without any AU annotation. We leverage prior probabilities on AUs, including expression-dependent AU probabilities and expression-independent AU probabilities, which are obtained from facial anatomy and emotion studies, and are independent of datasets. Unlike existing works which use AU labels for AU classifier learning, we adopt the probabilities on AUs to train AU classifiers. The generic knowledge from facial anatomy and long-term expert observations and studies is almost universally applicable to different people in real applications. The knowledge is imposed as soft probabilistic constraints to train the AU classifiers.

The main contributions of this paper are summarized as follows. Firstly, we propose a knowledge-driven method to jointly learn multiple AU classifiers by leveraging probabilities on AUs instead of AU annotations. We systematically summarize different types of generic knowledge from facial anatomy and emotion studies, including a variety of single and joint AU probabilities, and formulate AU recognition as a joint classifier and label learning problem. Secondly, we propose an algorithm to optimize the formulated problem by iteratively updating classifiers and AU labels. Thirdly, we evaluate the proposed method on five benchmark databases and compare it to the state-of-the-art methods.

*Corresponding author.

2. Related Work

AU recognition is a multi-label classification problem, since multiple AUs may be present simultaneously. Recently, several works exploit AU relations to facilitate the learning process of AU classifiers through either generative or discriminative strategy. Generative models are used to model structural outputs such as [10, 17, 26]. Tong *et al.* [20] proposed Bayesian Network (BN) to capture the dependencies among multiple AUs through its structure and conditional probabilities. Wang *et al.* [25] used Restricted Boltzman Machine (RBM) to capture dependencies among AUs. Unlike the above two works which learn AU dependencies from ground truth AU labels, Li *et al.* [11] proposed to learn a BN from the pseudo-data which is generated according to AU dependencies summarized from prior knowledge. Similar idea was also used in [4]. These methods still require ground-truth AU labels to obtain AU measurement.

For discriminative approaches, the dependencies among AUs are embodied by introducing the constraints into the objective function. Zhao *et al.* [32] used the constraints of group sparsity and positive and negative AU correlations to learn multiple AU classifiers. EmotionNet [2] is a supervised learning method for AU prediction. It replaces human AU annotation with automatic annotation by AU recognition algorithms. It complements our approach in the sense that we use generic knowledge to replace human annotation. Eleftheriadis *et al.* [7] proposed a multi-conditional latent variable model to integrate AU label dependencies into a latent space and learn AU classifiers. Zhang *et al.* [30] proposed to use multi-task multiple kernel learning to detect multiple AUs simultaneously by leveraging their intrinsic relations. Almaev *et al.* [1] constructed person-specific models with considering both relations across subjects and AUs. Zhao *et al.* [34] proposed a patch-based CNN model for region learning and AU detection. However, all of them require complete AU annotations for training. The models trained on one dataset are limited to the dataset where they are trained and cannot generalize well among datasets. A few works consider AU classification under partial AU annotations such as [27, 23, 28, 12, 24, 29].

To the best of our knowledge, only one work can handle AU recognition without AU annotation. Ruiz *et al.* [19] proposed to learn AU classifiers with expression labels, but without AU labels. Their model consists of two layers. The hidden layer is the AU classifiers and the visible layer maps the AU labels to expression. They use expression-dependent single AU probabilities to generate AU labels and use the generated samples to learn the visible layer by training a linear classifier. Each AU label is generated independently. Then, given the image and expression pairs, they train the hidden layer. The method has the following drawbacks. Firstly, it only uses knowledge on individual AUs and ignores dependencies among AUs. Such knowledge may generate unreasonable AU labels, since AUs are dependent on each other due to either underlying facial anatomy or the need to produce meaningful facial expression. Without considering relationships among AUs, the generated unreasonable AU labels affect the learning of the visible layer, and further the learning error will propagate to the learning of the hidden layer. Secondly, the method requires exact probabilities for single AUs given the expression. But, in general case, the prior probabilities might be represented by inequalities rather than exact probabilities and some singe AU probabilities are even not available. Finally, their method, requires, expression level labels for every training sample if no AU labels are provided.

Unlike Ruiz *et al.*'s work, our method exploits not only individual AU probabilities but also joint AU probabilities that represent relationships among AUs and expression. For individual AU probabilities, we also include more types than [19]. Our model can handle both equality constraints for exact probabilities as well as inequality constraints for inexact probabilities. For joint AU probabilities, besides expression-dependent AU relationships, we also exploit expression-independent AU relationships, including relationships among AUs that are controlled by the underlying facial anatomy, independent of facial expression. As a result, by exploiting expression-independent AU relationships, our model can also use samples without expression labels during training. Finally, we introduce an analytical method to systematically incorporate the AU probabilities for simultaneous AU classifier learning and AU label learning. In summary, our method represents a significant theoretical extension to [19], and the extension leads to significant performance improvement over [19].

The rest is arranged as follows. We first identify related generic domain knowledge from existing theories or studies (Sec. 3), then represent the knowledge as probabilistic constraints on AU states and relationships (Sec. 4.1), and finally use the constraints for learning (Sec. 4.2).

3. AU Probabilities

Existing methods require AU annotations for classifier learning. Accurate AU annotation requires expertise and time. We propose to learn classifiers for multiple AUs by leveraging prior probabilities on AUs instead of knowing AU annotations. Prior probabilities can be categorized into two groups, i.e., expression-independent and expression-dependent AU probabilities. They represent the dependencies among AUs. The former is applied to all the samples including samples without expression while the latter is applied to samples with expression labels.

3.1. Expression-independent joint AU probabilities

As defined in FACS [6], AUs are used to capture human facial movements by their facial appearance. Each AU

Table 1. The expression-independent AU dependencies according to facial anatomy and FACS [6]

AU relations	AU pairs
Positive correlation	(1,2), (4,7) , (4,9), (7,9), (6,12), (9,17), (15,17), (15,24), (17,24), (23,24)
Negative correlation	(2,6), (2,7), (12,15), (12,17)

movement is controlled by one or a group of facial muscles. Because of the underlying facial anatomy, AUs may depend on each other, independent of facial expression. According to facial anatomy and FACS, some AUs are likely to co-occur while other AUs seldomly appear together. Since AUs are controlled by facial muscles, their dependencies depend on the underlying facial anatomy. The dependencies among AUs caused by the underlying mechanism of muscles, are called expression-independent joint AU dependencies. The dependencies can be divided into two parts, *i.e.*, positive and negative correlations. The positive correlation is that some AUs are likely to appear simultaneously because these AUs are controlled by the same muscle group or neighboring muscle groups. For example, both AU1 (inner brow raiser) and AU2 (outer brow raiser) are controlled by the muscle group of *Occipito fraontalis*. The contraction of the lateral part produces AU2, and the contraction of the central part produces AU1. Since the contraction of the lateral part always happens together with the contraction of the central part, the occurrence of AU2 increases the chance of the occurrence of AU1. Therefore, AU1 and AU2 are positively correlated.

On the other hand, the negative correlation is that some AUs never or seldom appear simultaneously. Due to the underlying facial anatomy, certain muscles can not activate simultaneously. As a result, the occurrence of some AU decreases the chance of the occurrence of another AU. For example, AU12 (lip corner puller) produced by the muscle group of *Zygomaticus Major* and AU15 (lip corner depressor) produced by the muscle group of *depressor anguli oris* are unlikely to appear simultaneously. Since the contraction of Zygomaticus Major inhibits the movement of depressor anguli oris, the occurrence of AU12 decreases the probability of the occurrence of AU15. Therefore, AU12 and AU15 are negatively correlated. The positive and negative AU correlations are summarized in Table 1. These relationships can be applied to all the training samples, including samples without expression labels.

3.2. Expression-dependent AU probabilities

3.2.1 Expression-dependent single AU probabilities

According to FACS [6], for each expression, AUs can be classified into *primary*, *secondary*, and *other* AUs depending on their roles in producing the expression (see Table 2). Primary AUs are the most important emotional AUs that can be clearly classified as or are strongly pertinent to one of the basic expressions. Secondary AUs may co-occur with primary AUs and they provide supplementary support for the expression. Firstly, under a certain expression, the chance of the occurrence of its primary AUs is larger than the chance of the absence. Secondly, for AUs that are neither primary nor secondary AUs, their chance to appear is less than their chance of not to appear.

To quantitatively reveal the relationships among expression and AUs, Du *et al.* [5] studied facial expressions, including 6 basic expressions (anger, disgust, fear, happiness, sadness and surprise) and compound expressions. Their study reports the probabilities for prototypical AUs under each basic expression and compound emotion category. The probabilities of prototypical AUs under 6 basic expressions are shown in Table 3. These probabilities can also be used to provide weak supervisory information for AU classifier learning if these probabilities are available.

3.2.2 Expression-dependent joint AU probabilities

AUs often appear together to create meaningful and natural expressions. We can extract expression-dependent probabilities on multiple AUs according to FACS in Table 2. For any expression, its primary AUs are more likely to appear than its secondary AUs. And its secondary AUs have larger chance to appear than its other AUs. For example, AU4 is a primary AU for anger, AU17 is a secondary AU, and AU15 is neither a primary or secondary AU. Therefore, under anger expression, AU7 is more likely to appear than AU17. And AU17 is more likely to appear than AU15.

Besides, to reveal the dependencies between combinations of AUs and expression, Wallace *et al.* [9] built the Emotional Facial Action Coding System (EMFACS) considering only emotion-related facial actions. The study reported the most frequent AU combinations under each of the 6 basic expressions (see Table 4). For example, AU1 (inner brow raiser) and AU2 (outer brow raiser) are in the upper face while AU26 (jaw drop) is in the lower face. Though they are not correlated just according to facial anatomy, they are likely to appear simultaneously under surprise. Since these dependencies represent the co-occurrence of AUs, they represent positive correlations.

4. Proposed Method

Notation Let $\mathcal{D} = \{\mathbf{x}_n, z_n\}_{n=1}^{N}$ denote the training set. $\mathbf{x}_n \in \mathbf{R}^d$ is a d-dimensional feature vector. $z_n \in \{1, 2, ..., K\}$ is the expression label. K is the number of expressions. $\mathbf{Z}$ are the expressions of samples in $\mathcal{D}$. $\mathbf{y}_n = \{y_n^m\}_{m=1}^{M}$ denotes the AU labels of the n-th sample, which are unknown. M is the number of AU labels. Let $\tilde{\mathbf{Y}} \in \{1, -1\}^{N \times M}$ denote the estimated AU labels for all training samples. $\tilde{\mathbf{Y}}^m$ are the m-th estimated AU labels of all training samples, while $\tilde{\mathbf{Y}}_k^m$ are the m-th esti-

Table 2. Expression-AU relationships according to FACS [6]. Mark 'A' and 'B' refer to primary and second AU respectively.

AU	1	2	4	5	6	7	9	10	12	15	16	17	20	23	24	25	26
anger			A	A		B						B		A	A		
disgust							A	A				B				B	
fear	A	A	A	A		A							A	B		B	
happiness					A	B			A							B	
sadness	A		B		B	A				A		B					
surprise	A	A		A							B					A	A

Table 3. The prior probabilities on single AUs adapted from [5]

Expression	Probabilities of prototypical AUs
anger	AU4 ($\geq$ 0.7), AU7 ($\geq$ 0.7), AU10 (0.26), AU17 (0.52), AU23 (0.29), AU24 ($\geq$ 0.7)
disgust	AU4 (0.31), AU9 ($\geq$ 0.7), AU10 ($\geq$ 0.7), AU17 ($\geq$ 0.7)
fear	AU1 ($\geq$ 0.7), AU2 (0.57), AU4 ($\geq$ 0.7), AU5 (0.63), AU20 ($\geq$ 0.7), AU25 ($\geq$ 0.7), AU26 (0.33)
happiness	AU6 (0.51), AU12 ($\geq$ 0.7), AU25 ($\geq$ 0.7)
sadness	AU1 (0.6), AU4 ($\geq$ 0.7), AU6 (0.5), AU15 ($\geq$ 0.7), AU17 (0.67)
surprise	AU1 ($\geq$ 0.7) , AU2 ($\geq$ 0.7), AU5 (0.66), AU25 ($\geq$ 0.7), AU26 ($\geq$ 0.7)

mated AU labels of samples with the k-th expression. Let $\mathbf{P}$ denote the prior probabilities. Part of the prior probabilities are not available. Let $p_k^m = P(y^m = 1|z = k)$ denote the prior probability for the m-th AU under the k-th expression. And $\tilde{p}_k^m$ denotes its estimated probability from $\tilde{\mathbf{Y}}$. Let P^{i_1} denote the estimated marginal probability $P(y^i = 1)$, and $P_k^{i_1}$ denote the estimated conditional probability $P(y^i = 1|z = k)$. Let $P^{i_1 j_1}$ denote the estimated joint probability of co-occurrence of the i-th and the j-th AUs, *i.e.*, $P^{i_1 j_1} = P(y^i = 1, y^j = 1)$. Similarly, $P^{i_0 j_1} = P(y^i = 0, y^j = 1)$. $P^{i_1 j_1}$ and $P^{i_0 j_1}$ are also computed from $\tilde{\mathbf{Y}}$. Let $P_k^{i_1 j_1}$ denote the conditional joint probability $P(y^i = 1, y^j = 1|z = k)$. Let $S_{\mathcal{P}}$ and $S_{\mathcal{N}}$ denote the sets of expression-independent positively and negatively correlated AU pairs respectively. $S = S_{\mathcal{P}} \cup S_{\mathcal{N}}$. Let $E_{\mathcal{P}}$ denote the set of expression-dependent correlated AU pairs. And $E_{\mathcal{P}}^k \subset E_{\mathcal{P}}$ is for the k-th expression. For expression z, AUs are divided into three groups according to FACS, including primary AUs (y^p), secondary AUs (y^s), and other AUs (y^o). Let $[\cdot]_+$ represent $[t]_+ = \max(0, t)$. $|t|$ represents the absolute value of t.

4.1. Representation of AU probabilities

Expression-independent joint AU probabilities The expression-independent joint AU probabilities are applied to AUs of all training samples. We consider pairwise dependencies, including positive correlation and negative correlation between two AUs. The positive correlation can be interpreted in two ways, *i.e.*,

$$P(y^i = 1|y^j = 1) > P(y^i = 0|y^j = 1), \quad (1)$$

$$P(y^i = 1|y^j = 1) > P(y^i = 1|y^j = 0). \quad (2)$$

Table 4. Expression-dependent AU dependencies adapted from EMFACS [9]

Expression	AUs
anger	4+5, 4+7, 4+5+7, 17+24, 23
disgust	9, 10
fear	1+2+4, 20
happiness	12, 6+12, 7+12
sadness	1, 1+4, 15, 6+15, 11+15, 11+17
surprise	1+2+5AB, 1+2+26, 1+2+5AB+26

The first indicates that when one AU appears, the other AU is more likely to appear than not. The second indicates that the chance of the occurrence of one AU when the other appears is higher than when the other does not appear. From Eq. 2, we have $\frac{P(y_j=1|y_i=1)P(y_i=1)}{P(y_j=1)} > \frac{P(y_j=0|y_i=1)P(y_i=1)}{P(y_j=0)}$, then $\frac{P(y_j=1|y_i=1)P(y_i=1)}{P(y_j=1)} > \frac{[1-P(y_j=1|y_i=1)]P(y_i=1)}{[1-P(y_j=1)]}$. We get $P(y_j = 1|y_i = 1) > P(y_j = 1)$. The equivalent representations are

$$P(y^i = 1, y^j = 1) > P(y^i = 0, y^j = 1), \quad (3)$$

$$P(y^i = 1, y^j = 1) > P(y^i = 1)P(y^j = 1). \quad (4)$$

Similarly, for the negative correlation, we have the following two inequalities, i.e.,

$$P(y^i = 1, y^j = 1) < P(y^i = 0, y^j = 1), \quad (5)$$

$$P(y^i = 1, y^j = 1) < P(y^i = 1)P(y^j = 1). \quad (6)$$

Constraints on AU pairs enable the joint learning of multiple AU classifiers instead of learning classifiers individually. We define the loss for a correlated AU pair as follows

$$\ell_c(\tilde{\mathbf{Y}}^i, \tilde{\mathbf{Y}}^j) = \begin{cases} [P^{i_1}P^{j_1} - P^{i_1 j_1}]_+ + [P^{i_0}P^{j_1} - P^{i_1 j_1}]_+ \\ +[P^{i_1}P^{j_0} - P^{i_1 j_1}]_+, \forall (i,j) \in S_{\mathcal{P}} \\ [P^{i_1 j_1} - P^{i_1}P^{j_1}]_+ + [P^{i_1 j_1} - P^{i_0}P^{j_1}]_+ \\ +[P^{i_1 j_1} - P^{i_1}P^{j_0}]_+, \forall (i,j) \in S_{\mathcal{N}} \end{cases}, \quad (7)$$

where P^{i_1}, P^{i_0}, $P^{i_1 j_1}$, and $P^{i_1 j_0}$ are computed by using the estimated AU labels. They are computed as $P^{i_1} = \frac{\sum_n \delta(\tilde{y}_n^i = 1)}{N}$ and $P^{i_0 j_1} = \frac{\sum_n \delta(\tilde{y}_n^i = 0)\delta(\tilde{y}_n^j = 1)}{N}$. Each term represents a constraint of joint AU dependencies. The total loss of expression-independent joint AU probabilities is

$$L_c(\tilde{\mathbf{Y}}) = \sum_{(i,j) \in S} \ell_c(\tilde{\mathbf{Y}}^i, \tilde{\mathbf{Y}}^j). \quad (8)$$

Expression-dependent single AU probabilities According the specification of primary and secondary AUs in FACS (Table 2), we extract expression-dependent probabilities. If

an AU is a primary AU for expression z, the probability of its occurrence should be larger than the probability of its absence, *i.e.*, $P(y^p = 1|z) > P(y^p = 0|z)$. When $z = k$, the loss is defined as

$$\ell_s^{A_1}(\tilde{p}_k^i) = [0.5 - \tilde{p}_k^i]_+, i \in A_1^k, \quad (9)$$

where $\tilde{p}_k^i$ is the estimated probability of the primary AUi under expression k. A_1^k is the set of primary AUs. $\tilde{p}_k^i$ is computed as $\tilde{p}_k^i = \frac{\sum_n \delta(z_n=k)\delta(\tilde{y}_n^i=1)}{\sum_n \delta(z_n=k)}$, where $\delta(t) = 1$ if t is true. If the AU is neither a primary AU nor a secondary AU, the probability of its occurrence should be less than the probability of its absence, *i.e.*, $P(y^o = 1|z) < P(y^o = 0|z)$. When $z = k$, the loss is defined as

$$\ell_s^{A_2}(\tilde{p}_k^i) = [\tilde{p}_k^i - 0.5]_+, i \in A_2^k, \quad (10)$$

where $\tilde{p}_k^i$ is its estimated probability under expression k. A_2^k is the set of such AUs under this expression.

For primary AUs and AUs that are neither primary nor secondary AUs, the total loss of their probabilities is

$$L_s^A(\tilde{\mathbf{Y}}, \mathbf{Z}) = \sum_{k=1}^{K} \left(\sum_{i \in A_1^k} \ell_s^{A_1}(\tilde{p}_k^i) + \sum_{i \in A_2^k} \ell_s^{A_2}(\tilde{p}_k^i) \right). \quad (11)$$

In addition, we can extract expression-dependent single AU probabilities according to Table 3 if these probabilities are available. The probabilities of AUs given expression are represented by $P(y^i = 1|z) = \alpha$ or $P(y^i = 1|z) \geq \alpha$. The specific values of α for AUs are listed in the table. When $z = k$, the loss is defined as

$$\ell_s(\tilde{p}_k^i, p_k^i, s_k^i) = \begin{cases} [\tilde{p}_k^i - p_k^i]_+, s_k^i = -1 \\ |\tilde{p}_k^i - p_k^i|, s_k^i = 0 \\ [p_k^i - \tilde{p}_k^i]_+, s_k^i = 1 \end{cases}, i \in O^k, \quad (12)$$

where O^k is the set of AUs that have probabilities under expression k. $s_k^i = -1, 0$, or 1 represent that the probability is less, equal, or larger than certain value. $\mathbf{S}$ is the set of s_k^i. The total loss of these probabilities is

$$L_s(\tilde{\mathbf{Y}}, \mathbf{P}, \mathbf{S}, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{i \in O^k} \ell_s(\tilde{p}_k^i, p_k^i, s_k^i). \quad (13)$$

It can be dropped if the single AU probabilities in Table 3 are not available or cannot generalize across databases.

Expression-dependent joint AU probabilities The expression-dependent joint probabilities are applied to a group of samples with the same expression. We can extract two types of expression-dependent probabilities on multiple AUs according to FACS. Firstly, the probabilities of primary AUs should be larger than secondary AUs, *i.e.*, $P(y^p = 1|z) > P(y^s = 1|z)$. When $z = k$, the loss is

$$\ell_c^{B_1}(\tilde{p}_k^i, \tilde{p}_k^j) = [\tilde{p}_k^j - \tilde{p}_k^i]_+, (i,j) \in B_1^k, \quad (14)$$

where $\tilde{p}_k^i$ is the estimated probability of the primary AUi under expression k. $\tilde{p}_k^j$ is the estimated probability of the secondary AUj. B_1^k is the set of primary and secondary AU pairs under expression k. Secondly, the probabilities of secondary AUs should be larger than other AUs, *i.e.*, $P(y^s = 1|z) > P(y^o = 1|z)$. The loss is defined as

$$\ell_c^{B_2}(\tilde{p}_k^i, \tilde{p}_k^j) = [\tilde{p}_k^j - \tilde{p}_k^i]_+, (i,j) \in B_2^k, \quad (15)$$

where $\tilde{p}_k^i$ is the estimated probability of the secondary AUi under expression k. $\tilde{p}_k^j$ is the estimated probability of other AUj. B_2^k is the set of secondary and other AU pairs under expression k. The total loss of these expression-dependent joint AU probabilities is

$$L_c^B(\tilde{\mathbf{Y}}, \mathbf{Z}) = \sum_{k=1}^{K} \left(\sum_{(i,j) \in B_1^k} \ell_c^{B_1}(\tilde{p}_k^i, \tilde{p}_k^j) + \sum_{(i,j) \in B_2^k} \ell_c^{B_2}(\tilde{p}_k^i, \tilde{p}_k^j) \right).$$

Besides, we can also leverage expression-dependent correlated AU pairs in Table 4. The positive correlation represented as follows

$$P(y^i = 1, y^j = 1|z) > P(y^i = 0, y^j = 1|z), \quad (16)$$

$$P(y^i = 1, y^j = 1|z) > P(y^i = 1)P(y^j = 1|z). \quad (17)$$

We define the loss of expression-dependent correlated AU pairs when $z = k$ as follows

$$\ell_c^E(\tilde{\mathbf{Y}}_k^i, \tilde{\mathbf{Y}}_k^j) = [P_k^{i_1} P_k^{j_1} - P_k^{i_1 j_1}]_+ + [P_k^{i_0} P_k^{j_1} - P_k^{i_1 j_1}]_+ + [P_k^{i_1} P_k^{j_0} - P_k^{i_1 j_1}]_+, \forall (i,j) \in E_{\mathcal{P}}^k. \quad (18)$$

$P_k^{i_1}$ and $P_k^{i_1 j_1}$ are computed as $P_k^{i_1} = \frac{\sum_n \delta(z_n=k)\delta(\tilde{y}_n^i=1)}{\sum_n \delta(z_n=k)}$, $P_k^{i_1 j_1} = \frac{\sum_n \delta(z_n=k)\delta(\tilde{y}_n^i=1)\delta(\tilde{y}_n^j=1)}{\sum_n \delta(z_n=k)}$. The total loss of these expression-dependent joint AU probabilities is

$$L_c^E(\tilde{\mathbf{Y}}, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{(i,j) \in E_{\mathcal{P}}^k} \ell_c^E(\tilde{\mathbf{Y}}_k^i, \tilde{\mathbf{Y}}_k^j). \quad (19)$$

4.2. Formulation

The classification loss is defined as $L(\tilde{\mathbf{Y}}, \mathbf{X}; \mathbf{W}) = \frac{1}{NM} \sum_{m=1}^{M} \sum_{n=1}^{N} \ell(\tilde{y}_n^m, \mathbf{x}_n; \mathbf{W}^m)$. $\mathbf{W}^m$, the parameter of the m-th classifier, is the m-th column of $\mathbf{W}$. $\ell(\tilde{y}_n^m, \mathbf{x}_n; \mathbf{W}^m)$ is the loss of classifiers. It can be $\ell(y, \mathbf{x}; \mathbf{w}) = [1 - y(\mathbf{w}^T \mathbf{x})]_+$ for hinge loss and $\ell(y, \mathbf{x}; \mathbf{w}) = \log(1 + e^{-y\mathbf{w}^T \mathbf{x}})$ for logistic loss.

Given the formulations of AU probabilities, we now introduce our method to learn AU classifiers subject to these AU probabilities. We propose to jointly learn both multiple AU classifiers and AU labels of training samples by leveraging prior knowledge as follows

$$\begin{aligned} \min_{\tilde{\mathbf{Y}}, \mathbf{W}} \quad & L(\tilde{\mathbf{Y}}, \mathbf{X}; \mathbf{W}) + \lambda_c L_c(\tilde{\mathbf{Y}}) \\ & + \lambda_s^A L_s^A(\tilde{\mathbf{Y}}, \mathbf{Z}) + \lambda_s L_s(\tilde{\mathbf{Y}}, \mathbf{P}, \mathbf{S}, \mathbf{Z}) \\ & + \lambda_c^E L_c^E(\tilde{\mathbf{Y}}, \mathbf{Z}) + \lambda_c^B L_c^B(\tilde{\mathbf{Y}}, \mathbf{Z}) \\ s.t. \quad & \tilde{y}_n^m \in \{-1, 1\}, \\ & n \in \{1, ..., N\}, m \in \{1, ..., M\}, \end{aligned} \quad (20)$$

where λ_c, λ_s^A, λ_s, λ_c^E, and λ_c^B are the penalty factors. The first term is the classification loss. The second term is the loss of expression-independent joint AU probabilities from

Table 1. The third term is the loss of probabilities of single AUs from Table 2. The fourth term is the loss of probabilities of single AUs from Table 3, which can be removed if the probabilities are not available. The last two terms are loss of expression-dependent joint AU probabilities from Table 2 and 4. The knowledge is encoded as soft constraints which can be violated. Please note Eq. 20 can use both samples with expression labels as well as samples without expression labels since the second term can be applied to all the samples. As a result, our method technically can be solved without even expression labels, as we can only use the second term in Eq. 20. During testing, the learned AU classifiers are applied individually. No expression and AU relationships are needed.

4.3. Optimization

We propose an iterative algorithm under the alternating optimization framework (see Algo. 1). Given $\tilde{\mathbf{Y}}$, it becomes a convex problem with respect to $\mathbf{W}$. Given $\mathbf{W}$, we use a greedy strategy to estimate $\tilde{\mathbf{Y}}$. We use the knowledge from Table 2 to initialize $\tilde{\mathbf{Y}}$. For a sample with expression, we assign 1 to the primary or secondary AUs and 0 to other AUs. Then, we use the assigned labels to initialize $\mathbf{W}$.

Fix $\tilde{\mathbf{Y}}$, optimize $\mathbf{W}$ Given $\tilde{\mathbf{Y}}$, the parameters of each classifier are not coupled, so each classifier can be learned separately by solving the following subproblem

$$\min_{\mathbf{W}^m} \frac{1}{N}\sum_{n=1}^{N} \ell(\tilde{y}_n^m, \mathbf{x}_n; \mathbf{W}^m). \tag{21}$$

The subproblem can be solved efficiently by existing optimization algorithms. We use LBFGS [16] for optimization.

Fix $\mathbf{W}$, optimize $\tilde{\mathbf{Y}}$ We use a greedy strategy to minimize the objective function iteratively. An iteration consists of three steps as shown in Algo. 1. We can find the best AU configuration for each sample by solving the subproblem

$$\begin{aligned}\min_{\tilde{\mathbf{y}}_n} \ & \frac{1}{M}\sum_{m=1}^{M} \ell(\tilde{y}_n^m, \mathbf{x}_n; \mathbf{W}^m) + \lambda_c L_c(\tilde{\mathbf{Y}}) \\ & + \lambda_s^A L_s^A(\tilde{\mathbf{Y}}, \mathbf{Z}) + \lambda_s L_s(\tilde{\mathbf{Y}}, \mathbf{P}, \mathbf{S}, \mathbf{Z}) \\ & + \lambda_c^E L_c^E(\tilde{\mathbf{Y}}, \mathbf{Z}) + \lambda_c^B L_c^B(\tilde{\mathbf{Y}}, \mathbf{Z}) \\ s.t. \ & \tilde{y}_n^m \in \{-1, 1\}, m \in \{1, ..., M\}. \end{aligned} \tag{22}$$

The first step for each sample is independent and the evaluation of each configuration is also independent. Therefore, it can be implemented parallelly. The alternating optimization procedure is guaranteed to converge because the objective is minimized at each step and it is non-increasing. The complexity of finding the best $\tilde{\mathbf{Y}}$ by brute-force is $\mathcal{O}(2^{NM})$. Our greedy approach reduces the complexity to $\mathcal{O}(N^2 2^M)$.

5. Experiments

The goal of experiments is to show that our weakly supervised method can achieve better performance than the competing methods and achieve comparable performance to the methods that use AU annotations.

Algorithm 1 Model learning with AU probabilities

Input: training data $\mathcal{D}$,
prior probabilities on single AUs $P(y^i|z)$,
expression-independent AU pairs S, and
expression-dependent AU pairs $E_{\mathcal{P}}$.
Output: $\mathbf{W}$, $\tilde{\mathbf{Y}}$

Initialize $\tilde{\mathbf{Y}}$ with single probabilities and update $\mathbf{W}$
while not converging **do**
  Fix $\mathbf{W}$, update the labels $\tilde{\mathbf{Y}}$
    Step 1: find the best configuration for each sample by solving Eq. 22
    Step 2: (a) compute the objective under the best configuration of each sample (Eq. 20)
    (b) compare the objective values of samples
    (c) select the sample with the minimum objective
    Step 3: update only the selected sample with its best configuration and keep the previous configurations of other samples
  Repeat Step 1∼3 until no reduction of the objective
  Fix $\tilde{\mathbf{Y}}$, update $\mathbf{W}$ by solving Eq. 21
end while
return $\mathbf{W}$, $\tilde{\mathbf{Y}}$

5.1. Settings

Datasets: The CK+ database [13] is a posed expression database. Apex frames from 309 sequences of 109 subjects with 6 basic expressions are collected. The MMI database [21] is a posed expression database. Apex frames from 196 sequences of 27 subjects with both AUs and expression are collected. The BP4D database [31] is a spontaneous expression database. 391 apex frames from 41 subjects with 6 basic expressions are extracted. The EmotioNet database (ENet) [8] is collected in the wild. Few images have both basic expressions and AU annotations. In our experiments, 345 images with only basic expression annotations and 420 images with only AU annotations are used, called as ENet-E and ENet-AU respectively. We consider 8 AUs in CK+, MMI and BP4D and consider 6 AUs in ENet. We use the prior probabilities related to these AUs from Tables 1, 2, 3, and 4.

Features: We use [15] to detect facial landmarks. Face images are aligned according to the two eye centers. We extract LBP (Local Binary Pattern) [18] features around 51 inner landmarks. The patch size is 32×32. To evaluate our methods when using different features, we also use the coordinates of the 51 landmarks as another type of feature. If not specified, the LBP features and the hinge loss are used.

Evaluation metric: We use F1 score as evaluation metric, i.e., F1 $= \frac{2 \cdot R \cdot P}{R+P}$, where R is recall and P is precision.

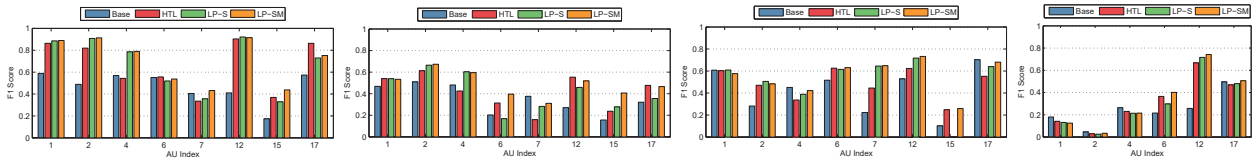

Figure 1. Within database performance for each AU. From left to right are the results on CK+, MMI, BP4D, and ENet.

The hyperparameters are selected through cross validation on the training set. We use the error between the estimated probabilities and the prior probabilities on single AUs as the measure for the selection. The ranges of hyperparameters are $\lambda_s, \lambda_s^A, \lambda_c, \lambda_c^e, \lambda_c^B \in \{10, 10^2, 10^3, 10^4\}$.

Comparison: Firstly, we perform within-database AU recognition using different features and different losses for classifiers. We compare our method (LP-SM) to weakly supervised methods such as Base, LP-S, and HTL [19]. **Base** is the method that exploits the prior probabilities to assign AU labels to samples by using the threshold as 0.5. The classifiers are trained on the assigned labels. **HTL** is currently the only state-of-the-art method that learns AU classifiers without AU annotations. It can handle only single AU probabilities. **LP-SM** uses both single and joint AU probabilities. **LP-S** is a variant of LP-SM, which uses only single AU probabilities. We use 5 fold subject independent cross validation for evaluation. Secondly, we compare to the state-of-the-art supervised learning methods such as **JPML** [32], **SVM-HMM** [22], **HRBM** [25], and **MC-LVM** [7]. They use AU annotations to train classifiers.

5.2. Results

5.2.1 Evaluation of our method

Learning with different features and classifiers. We evaluate the performance of our method with different image features and different losses for classifiers. Table 5 shows the performance of using different features, i.e., LBP and geometry features. The classification loss is hinge loss for Base, LP-S, and LP-SM. Table 6 shows the performance of using different losses of classifiers, i.e., hinge loss and logistic loss. The features are LBP. Fig. 1 shows the detailed performance of each AU when using LBP features and hinge loss. For the ENet database, we use ENet-E for training and ENet-AU for testing.

Firstly, LP-SM outperforms Base, HTL, and LP-S on all the three databases when using different features or losses for classifiers. These results demonstrate that the generic knowledge on AUs is still applicable when using different features or losses. Compared to LP-S, LP-SM leverages not only the prior knowledge on single AUs, but also relationships among multiple AUs. Compared HTL, HTL learns the mapping from AU to expression first and then learn the AU classifiers. The learning error at the first stage affects the second stage. Unlike HTL, we jointly optimize the AU labels and classifiers to avoid error propagation. We also employ a larger set of AU probabilities to help learn the classifiers, including expression-independent and expression-dependent joint AU probabilities. Secondly, Table 7 shows the cross-database evaluation of SVM that uses AU labels for training on a source database and for testing on a target database. The performance of SVM drops dramatically. Compared to SVM, we apply the same generic knowledge to different databases and our method achieves better performance. This demonstrates that the generic knowledge can generalize to different databases.

Table 5. Performance of using different features

Database	LBP				Landmarks			
	Base	HTL [19]	LP-S	LP-SM	Base	HTL [19]	LP-S	LP-SM
CK	0.470	0.657	0.679	**0.719**	0.570	0.689	0.689	**0.732**
MMI	0.348	0.415	0.419	**0.508**	0.413	0.438	0.442	**0.481**
BP4D	0.427	0.488	0.515	**0.564**	0.435	0.470	0.504	**0.563**
ENet	0.243	0.317	0.311	**0.336**	0.202	0.311	0.319	**0.337**

Table 6. Performance of using different losses for classifiers

Database	hinge loss			logistic loss		
	Base	LP-S	LP-SM	Base	LP-S	LP-SM
CK	0.470	0.679	**0.719**	0.449	0.661	**0.690**
MMI	0.348	0.419	**0.508**	0.360	0.370	**0.481**
BP4D	0.427	0.515	**0.564**	0.433	0.501	**0.552**
ENet	0.243	0.311	**0.336**	0.239	0.325	**0.351**

Table 7. Cross-database evaluation of SVM.

	target		
source	CK	MMI	BP4D
CK	**0.783**	0.429	0.352
MMI	0.551	**0.519**	0.509
BP4D	0.403	0.431	**0.667**

Contribution of joint AU probabilities. To investigate the contribution of expression-independent and expression-dependent relationships, we drop one of them during learning. The results are shown in Table 8. **rmInd** represents learning AU classifiers by dropping expression-independent relationships while **rmDep** represents dropping expression-dependent relationships. The results show that both expression-independent and expression-dependent relationships contribute, but the expression-independent relationships are more important.

Table 8. Comparison of expression-independent and expression-dependent joint AU relationships

Method	rmInd	rmDep	LP-SM
CK+	0.686	0.700	**0.719**
MMI	0.414	0.452	**0.508**
BP4D	0.493	0.550	**0.564**
ENet	0.312	0.322	**0.336**

Comparison of the sources of single AU probabilities. The single AU probabilities come from two sources, i.e., the FACS [6] and the study [5]. To verify the informativeness of each source, we compare the performance of using single AU probabilities from only the FACS or the study [5].

FACS provides generic singe AU probabilities based on the specification of primary, secondary, and other AUs. The study [5] provides more informative probabilities for specific AUs under different expressions. The results are shown in Table 9. It shows that both sources of single probabilities contribute and the study [5] is more informative. It also shows to some extent that the probabilities in [5] can generalize across datasets.

Table 9. Comparison of sources of single AU probabilities

Source	FACS	Study [5]	FACS & Study [5]
CK+	0.680	0.708	**0.719**
MMI	0.464	0.488	**0.508**
BP4D	0.344	0.554	**0.564**
ENet	0.278	0.334	**0.336**

Learning individual classifiers. The purpose of multi-label learning is to use label correlations to improve individual classifiers. Though all AUs are trained jointly, during recognition, they are applied individually. To verify whether AU detectors are learned for individual AUs or their combinations, we analyze the predictions of classifiers on the CK+ database. The distributions of AU pairs in the prediction are shown in Table 10. The results show that our joint learning method generates individual classifiers rather than classifiers for AU combinations since P(A=1,B=0) or P(A=0,B=1) of predicted AU pairs are larger than 0.

Table 10. Co-occurrence of AU pairs in the prediction

(A,B)	Positive correlation				Negative correlation			
	(1,2)	(4,7)	(6,12)	(15,17)	(2,6)	(2,7)	(12,15)	(12,17)
P(A=1,B=1)	0.30	0.23	0.18	0.28	0.13	0.08	0.04	0.04
P(A=0,B=1)	0.08	0.09	0.09	0.03	0.37	0.26	0.15	0.08
P(A=1,B=0)	0.05	0.05	0.14	0.15	0.16	0.22	0.08	0.39

Semi-supervised Learning. We can extend our model to a semi-supervised model by adding a term for samples with AU annotations, *i.e.*, $\bar{L}(\mathbf{Y}, \mathbf{X}; \mathbf{W}) = \frac{1}{NM}\sum_{m=1}^{K}\sum_{n=1}^{N}\ell(y_n^m, \mathbf{x}_n; \mathbf{W}^m)$. We perform an experiment on CK+ under the scenario that half of training samples have AU annotations. The F1 scores of using only AU annotations, using only prior probabilities, and using both are 0.724, 0.719, and 0.754 respectively. It future demonstrates the power of the prior probabilities.

5.2.2 Comparison to the state-of-the art methods

We compare to the supervised methods that require AU annotations. The results are shown in Table 11. (*) indicates reported results. Though the performance of our method is not as good as the fully supervised methods, surprisingly, it achieves promising comparable performance to them without using any AU annotations, especially on CK+ and MMI. Note that our method applies the same prior probabilities to different databases while other methods use the AU annotations in each database. Though they perform well in within-database evaluation, the performance drops in cross-database evaluation (see Table 7). The performance on ENet is not as good as CK+ since training images have low quality and contain large pose and illumination variance.

Table 11. Comparison to the state-of-the-art supervised methods.

Method	CK+	MMI	BP4D	ENet
JPML [32]	0.788*	-	0.676*	-
SVM-HMM [22]	-	**0.555***	-	-
HRBM [25]	0.792	0.547	**0.688**	**0.436**
MC-LVM [7]	**0.801***	-	-	-
LP-SM	0.719	0.508	0.564	0.336

5.2.3 Experiments on the PAIN database

To further evaluate the generalization ability, we apply our method to the UNBC-McMaster Shoulder Pain Expression Archive (PAIN) database [14]. The majority faces have no pain in PAIN. 300 apex frames under pain and 300 frames without pain are collected. We consider the recognition of AU6, AU7, and AU12 since other AUs rarely appear. We extract knowledge from the definition of pain , i.e. pain = AU4 + max(AU6,AU7) + max(AU9,AU10) + AU43. Firstly, for the non-pain expression, we have $P(\text{AU}_i = 1|\text{nopain}) < P(\text{AU}_i = 1|\text{pain})$ for $i = 6, 7$. Since AU6 and AU7 are the dominant AUs, we have $P(\text{AU}_i = 1|\text{pain}) > P(\text{AU}_i = 0|\text{pain})$ for $i = 6, 7$. We also have the expression-independent correlation, i.e., (AU6, AU12). The results are shown in Table 12. The supervised methods achieve better performance since they have the AU annotations which provide strong supervisory information. But, the performance of our method is much better than the random guess (F1=0.369). It still shows that the generic knowledge is applicable on the PAIN database and provides useful constraints for classifiers in the solution space.

Table 12. Performance on the PAIN database

AU	SVM [14]	MC-LVM [7]	LP-SM
6	0.774	0.987*	0.557
7	0.695	0.679*	0.521
12	0.844	-	0.457
Avg.	0.771	0.833*	0.512

6. Conclusion

In this paper, we propose to use generic domain knowledge to train AU classifiers without AU annotations. Represented as AU probabilities and derived from the underlying facial anatomy, the domain knowledge imposes generic constraints on AU dependencies and emotion studies. We propose to simultaneously learn the AU classifiers and AU labels of training samples. Evaluations on five databases show our methods achieve comparable performance to fully supervised methods, but with much better generalization capabilities. Besides AU recognition, the proposed method can be applied to other applications if domain knowledge is provided, such as object/attribute recognition.

Acknowledgments: The work was accomplished when the first author was visiting Rensselaer Polytechnic Institute (RPI), through a scholarship from China Scholarship Council. The support of CSC and RPI is greatly appreciated. This work was also supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61672520, 61620106003, and the Beijing Natural Science Foundation under Grant No. 4162056.

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 *ICCV*, 2015. 2

[2] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *ICSMC*, 2004. 1, 2

[3] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016. 1

[4] J. Chen, S. Nie, and Q. Ji. Data-free prior model for upper body pose estimation and tracking. *TIP*, 2013. 2

[5] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *PNAS*, 2014. 3, 4, 7, 8

[6] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 2, 3, 4, 7

[7] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *ICCV*, 2015. 1, 2, 7, 8

[8] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016. 6

[9] W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 1983. 3, 4

[10] Y. T. Lei Zhang and Q. Ji. Probabilistic graphical models and their applications in computer vision. In *Handbook of pattern recognition and computer vision*. 2010. 2

[11] Y. Li, J. Chen, Y. Zhao, and Q. Ji. Data-free prior model for facial action unit recognition. *TAC*, 2013. 2

[12] 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. *PR*, 2016. 1, 2

[13] 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 *CVPR Workshops*, 2010. 6

[14] 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 *FG Workshop*, 2011. 8

[15] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004. 6

[16] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 1993. 6

[17] S. Nowozin, C. H. Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 2011. 2

[18] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002. 6

[19] A. Ruiz, J. Van de Weijer, and X. Binefa. From emotions to action units with hidden and semi-hidden-task learning. In *ICCV*, 2015. 2, 7

[20] Y. Tong and Q. Ji. Learning bayesian networks with qualitative constraints. In *CVPR*, 2008. 1, 2

[21] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, Malta, May 2010. 6

[22] 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)*, 2012. 7, 8

[23] J. Wang, S. Wang, and Q. Ji. Facial action unit classification with hidden knowledge under incomplete annotation. In *ICMR*, 2015. 2

[24] S. Wang, Q. Gan, and Q. Ji. Expression-assisted facial action unit recognition under incomplete au annotation. *PR*, 2017. 2

[25] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, 2013. 1, 2, 7, 8

[26] B. Wu. *Incorporating Target Prior Knowledge into Structured Output Learning Models*. PhD thesis, Chinese Academy of Sciences, 2014. 2

[27] B. Wu, Z. Liu, S. Wang, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels. In *ICPR*, 2014. 2

[28] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *PR*, 2015. 2

[29] S. Wu, S. Wang, B. Pan, and Q. Ji. Deep facial action unit recognition from partially labeled data. In *ICCV*, 2017. 2

[30] 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 *WACV*, 2014. 2

[31] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *FG Workshops*, 2013. 6

[32] 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 *CVPR*, 2015. 1, 2, 7, 8

[33] 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 and holistic expression recognition. *TIP*, 2016. 1

[34] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, 2016. 2