Joint Patch and Multi-label Learning for Facial Action Unit and Holistic Expression Recognition

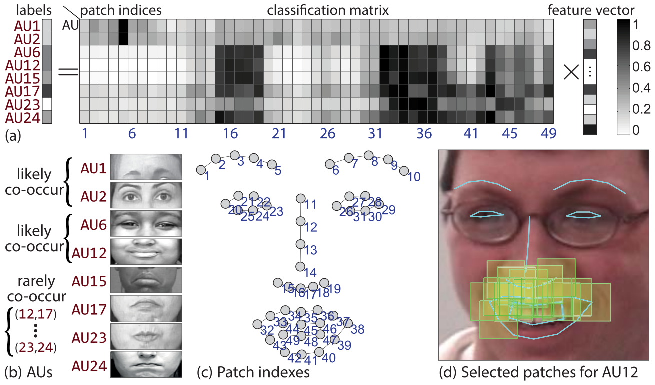
Kaili Zhao, Wen-Sheng Chu, *Student Member, IEEE*, Fernando De la Torre, Jeffrey F. Cohn, and Honggang Zhang, *Senior Member, IEEE*

***Abstract*— Most action unit (AU) detection methods use one-versus-all classifiers without considering dependences between features or AUs. In this paper, we introduce a joint patch and multi-label learning (JPML) framework that models the structured joint dependence behind features, AUs, and their interplay. In particular, JPML leverages group sparsity to identify important facial patches, and learns a multi-label classifier constrained by the likelihood of co-occurring AUs. To describe such likelihood, we derive two AU relations, *positive correlation* and *negative competition*, by statistically analyzing more than 350,000 video frames annotated with multiple AUs. To the best of our knowledge, this is the first work that jointly addresses patch learning and multi-label learning for AU detection. In addition, we show that JPML can be extended to recognize holistic expressions by learning common and specific patches, which afford a more compact representation than the standard expression recognition methods. We evaluate JPML on three benchmark datasets CK+, BP4D, and GFT, using within- and cross-dataset scenarios. In four of five experiments, JPML achieved the highest averaged F1 scores in comparison with baseline and alternative methods that use either patch learning or multi-label learning alone.**

***Index Terms*— Patch learning, multi-label learning, group sparsity, support vector machine, ADMM, facial expression recognition, facial action unit detection, correlation.**

I. Introduction

THE Facial Action Coding System (FACS) [11] is a comprehensive system for describing facial muscle movements. Anatomically-based actions, referred to as Action Units (AUs), alone and in thousands of combinations can account for nearly all possible facial expressions. However, this descriptive power is not without cost. Manual FACS coding is labor intensive. Training a classifier can require a hundred hours or more to reach acceptable competence. Once a FACS coder achieves this milestone, annotation (also referred to as coding) can require an hour or more for each 30 to 60 seconds of video, and inter-observer reliability must be closely monitored to maintain quality. To make possible more efficient use of FACS, computer vision strives for automatic AU coding. While significant progress has been made toward this goal [1], [6], [10], [26], at least two critical problems remain. These are patch learning and multi-label learning. *Patch learning (PL)* addresses how to effectively exploit local dependencies between features; *multi-label learning (ML)* seeks to exploit strong correlations between AUs.



(b) AUs (c) Patch indexes (d) Selected patches for AU12

Fig. 1. An illustration of the proposed Joint Patch and Multi-label Learning (JPML) framework: (a) Derivation of multiple AU labels through a learned group-sparsity-induced classification matrix and a given feature vector (*e.g.*, SIFT), (b) Likely and rarely co-occurring AUs, (c) Patch indices, and (d) Automatically selected patches. The example is for AU12 (lip corner puller). Details, *e.g.*, features, derivations, and the number of AUs, follow.

Most current AU detection approaches extract features across the entire face and concatenate them for learning. However, these approaches ignore the fact that faces are structured and thus neighboring features are dependent. As shown in Fig. 1(d), features extracted within a local region are correlated due to a shared neighborhood. In this paper, we define such regions as "patches". By jointly modeling features extracted within each patch, it is possible to identify informative regions as defined in FACS [11]. For instance, Zhong *et al.* [43] effectively applied patch learning to detect prototypic expressions (*e.g.*, happiness or sadness). Patch learning helps reduce feature dimensionality and promotes interpretability for the demanding AU detection problem.

Manuscript received August 6, 2015; revised February 6, 2016 and April 29, 2016; accepted April 30, 2016. Date of publication May 18, 2016; date of current version June 28, 2016. This work was supported in part by the National Science Foundation under Grant 1418520-L and Grant IIS 1418026, in part by the U.S. National Institutes of Health under Grant MH096951, and in part by the National Institute of Mental Health under Grant R21 MH099487-01A1. The work of K. Zhao was supported by the Beijing University of Posts and Telecommunications through the Excellent Ph.D. Students Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Clausi.

K. Zhao and H. Zhang are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: kailizhao@bupt.edu.cn; zhhg@bupt.edu.cn).

W.-S. Chu and F. De la Torre are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: wensheng.chu@gmail.com; ftorre@cs.cmu.edu).

J. F. Cohn is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA, and also with the Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: jeffcohn+@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2570550

Similarly, just as features within patches have dependencies, AUs are dependent as well. For instance, AU 1 (inner-brow raise) increases the likelihood of AU 2 (outer-brow raise) and decreases the likelihood of AU 6 (cheek raiser). Multi-label learning builds upon this knowledge. Using relations between AUs improves learning by implicitly increasing the sample size for correlated AUs. Recent efforts have explored AU relationships using Bayesian networks (BN) [32], [33] and Restricted Boltzmann Machine (RBM) [34]. Other works developed generic domain knowledge to learn AU models without training data [19].

Instead of modeling PL or ML individually, in this paper, we propose a novel framework, Joint Patch and Multi-label Learning (JPML), to address both problems with one stone. JPML aims to model the structured joint dependencies behind features, AUs, and their interplay. Specifically, JPML leverages the structure in the classification matrix and the multi-label nature in AU annotations, and naturally blends both tasks into one. To capture the dependencies between AUs, we explore two types of AU relations, termed *positive correlation* and *negative competition*, by statistically analyzing more than 350,000 annotated video frames from three datasets. The datasets involve two- and three-person social contexts and a variety of emotion inductions. Incorporating such AU relations, JPML is able to simultaneously select a discriminative subset of patches and learn a multi-AU classifier.

Fig. 1 illustrates the main idea of the proposed JPML: (a) shows that JPML learns a classification matrix for jointly predicting multiple AU labels and inducing group-sparsity among patches. The joint process of JPML are *mutually-beneficial* due to the complementary characteristics in the classification matrix. (b) shows likely and unlikely co-occurring AUs. (c) shows patch indices. (d) illustrates the patches selected by JPML, showing JPML's ability to find a discriminative subset of patches to identify AU12 (oblique lip corner puller). Comprehensive experiments on CK+ [22], GFT [27], and BP4D [40], validate the effectiveness of JPML in comparison to baseline and alternative approaches that use only patch learning or multi-label learning. To summarize, this paper presents three contributions:

- A novel Joint Patch and Multi-label Learning (JPML) framework that simultaneously leverages dependencies between features, AUs, and their interplay. To the best of our knowledge, this is the first study that jointly addresses patch and multi-label learning for AU detection.
- A derivation of AU relations by statistically analyzing over 350,000 annotated video frames coded with multiple AUs. We categorize the AU relations into *positive correlation* and *negative competition*, which coincide with findings in existing literature, including FACS.
- An extension of JPML that disentangles common and specific patches to recognize holistic facial expressions. In our experiments, this extension reveals sparse patches shared or specific to particular expressions, and achieves comparable accuracy with alternative approaches.

A preliminary version of this work appeared as [42]. In this paper, we provide technical details in solving the multi-label learning algorithm, present extended results on holistic expression recognition, and offer an in-depth analysis of what our models learned about important patches for each AU. The rest of this paper is organized as follows. Section II discusses related work. Section III presents the proposed JPML framework. Section IV introduces an extension of JPML that decomposes patches into common and specific ones for different expressions. Section V evaluates JPML against existing patch-learning and multi-label learning methods. Section VI concludes the paper with remarks and future work.

II. Related Work

Automatic facial AU detection has been a vital research domain to objectively describe facial actions related to emotion, intention, and related states. See [1], [6], [10], and [26] for comprehensive reviews. Our work closely follows recent efforts in patch learning and multi-label learning. Below, we review each in turn.

A. Patch Learning

AU detection methods often perform *feature learning* to select a representative subset of raw features. Standard feature learning approaches include AdaBoost [20], GentleBoost [14], and a linear SVM [22], which are often employed on feature descriptors such as DAISY/SIFT [45], LBP [16], or concatenated image features [31]. Because methods based on feature learning typically perform pooling from across the face, much valuable information could be lost. Based on FACS [11], AUs relate to different regions of the face. Different facial regions can provide unique information for discriminating facial AUs and holistic expressions [28], [29]. Cohn and Zlochower [7] found that facial regions differentially contribute to holistic expressions. Zafeiriou and Pitas [36] applied elastic graph matching to produce an expression-specific graph that identified the most discriminant facial landmarks for specific expressions. Specific facial regions are critical for detecting AUs. For example, if one seeks to detect AUs 1 and 2 (eye brow raiser), the regions of eyes and brows are likely to be more informative than the jaw.

Using domain knowledge, feature selection can be sampled within subregions, or *patches*, of the face. Following this intuition, *patch learning* was proposed to model the region specificity to improve the performance of AU detection. Zhong *et al.*. [43] divided a facial image into uniform patches and then categorized these patches into common ones and specific ones for holistic expression recognition. Following this idea, Liu *et al.* [21] proposed to select common and specific patches corresponding to an expression pair (*e.g.*, happiness-sadness). However, these patches were modeled implicitly and did not directly capture regional importance for facial AUs. Recently, Taheri *et al.* [30] used two-layer group sparse coding to encode AUs on predefined regions, and recovered facial expressions using sparsity in AU composition rules. Similarly, using predefined grids on faces, Kotsia and Pitas [17] designed an architecture to track grids for each AU and then adopted AU composition rules to predict holistic expressions.

These patch learning approaches have been proven effective when applied to posed expressions. However, spontaneous facial behavior often involve out-of-plane rotation that may introduce significant errors. In addition, it is unclear how AU dependencies can be taken into account.

B. Multi-Label Learning

Existing research suggests the existence of strong AU correlations [19], [34]. For instance, AUs 6 and 12 are known to co-occur in expressions of enjoyment, embarrassment, and pain but not in expressions of distress or sadness. Knowing such AU correlations could assist in predicting AUs given one another. Tong *et al.* [33] proposed to model AU correlations with Bayesian Network (BN) and initialized the network structure with inferred co-occurring relationships among AUs from FACS [11]. Based on the observed intra-, inter-, and rage-relationships between AUs, Tong and Ji *et al.* [32] constructed a BN with the corresponding relation constraints to help improve predictions in the limited training data scenario. These two works focus on the semantic local correlations between AUs. To this end, Wang *et al.* [34] used Restricted Boltzmann Machine (RBM) to capture the global correlations among AUs. Other approaches use generic domain knowledge. For example, AU correlations can be modeled as a directed graph without training data [19]. In addition, a sparse multi-task model can be employed to learn several AUs together [41]. Nevertheless, it is unclear how these methods can best identify a discriminative subset of patches to improve AU detection. This study proposes a Joint Patch and Multi-label Learning (JPML) framework that simultaneously addresses patch learning and multi-label learning for AU detection.

III. Joint Patch and Multi-Label Learning (JPML)

A. Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ be the training set with N instances and L AUs, where $\mathbf{x}_i \in \mathbb{R}^D$ is a feature vector from a facial image, and $\mathbf{y}_i \in \{+1, -1\}^L$ is an $L \times 1$ label vector which indicates a presence of the ℓ-th AU if the ℓ-th element $y_{i\ell} = +1$, and an absence of the ℓ-th AU if $y_{i\ell} = -1$ (see notation).[1] For brevity, we denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ as a data matrix, and $\mathcal{I}_\ell = \{i | y_{i\ell} = +1\}$ as an index set of instances that contain the ℓ-th AU. Our goal is to learn L linear classifiers in the matrix form $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L] \in \mathbb{R}^{D\times L}$ that enforces group-wise sparse feature selection (corresponding to the rows of $\mathbf{W}$) and AU relations (corresponding to the columns of $\mathbf{W}$). We formulate JPML as an unconstrained optimization:

$$\min_{\mathbf{W}\in\mathbb{R}^{D\times L}} \; L(\mathbf{W}, \mathcal{D}) + \alpha\Omega(\mathbf{W}) + \Psi(\mathbf{W}, \mathbf{X}), \tag{1}$$

where $L(\mathbf{W}, \mathcal{D}) = \sum_{\ell=1}^{L} \sum_{i\in\mathcal{I}_\ell} \ln(1 + \exp(-y_{i\ell}\mathbf{w}_\ell^\top \mathbf{x}_i))$ is the logistic loss, $\Omega(\mathbf{W})$ is the *patch regularizer* that enforces sparse rows of $\mathbf{W}$ as *groups*, and $\Psi(\mathbf{W}, \mathbf{X})$ is a *relational regularizer* that constrains predictions on $\mathbf{X}$ with AU relations. Tuning parameters are α for $\Omega(\cdot)$, and (β_1, β_2) that are included in $\Psi(\cdot, \cdot)$. Problem (1) involves two tasks: identify a discriminative subset of patches for each AU (*patch learning*), and incorporate AU relations into model learning (*multi-label learning*). Below we detail each task in turn.

B. Patch Learning

The first task of JPML addresses *patch learning*. According to FACS [11], AUs are muscle movements corresponding to particular facial regions. Unlike standard feature learning methods that treat features separately [20], [23], patch learning imposes local dependencies within facial patches. To this end, most existing works select patches on a uniformly distributed grid [21], [30], [43]; this paper exploits *landmark patches* that are centered at facial landmarks (as depicted in Fig. 1(c)). These landmark patches better capture the appearance changes at different facial regions (*e.g.*, mouth corner) because of the non-rigidity of faces. In particular, we describe each patch using a 128-D SIFT descriptor. Each facial image is then represented as a 6272-D feature vector by concatenating the SIFT descriptors over 49 facial landmarks.

To address the regional appearance changes on AUs, we define a group-wise sparsity on the classification matrix $\mathbf{W}$. Group sparsity learning aims to split variables into groups and then to select groups in sparsity. The effectiveness of group sparsity has been shown in recovering joint sparsity across input dimensions, *e.g.*, [18], [39]. Given the structural nature of our problem, within each column of $\mathbf{W}$, we split every 128 values into non-overlapping groups, where each group corresponds to the SIFT features extracted from a particular patch. This grouping encourages a sparse selection of patches by jointly setting a group of rows in $\mathbf{W}$ to zero. In particular, Problem (1) reduces to:

$$\min_{\mathbf{W}\in\mathbb{R}^{D\times L}} \; L(\mathbf{W}, \mathcal{D}) + \alpha\Omega(\mathbf{W}), \tag{2}$$

where $\Omega(\mathbf{W}) = \sum_{\ell=1}^{L} \sum_{p=1}^{49} \|\mathbf{w}_\ell^p\|_2$ is the *patch regularizer*, and $\mathbf{w}_\ell^p$ is the p-th group for the ℓ-th AU, *i.e.*, rows of $\mathbf{w}_\ell$ grouped by the patch p. In other words, the regularizer $\Omega(\mathbf{W})$ is an L_{21}-norm that induces sparsity in terms of “groups”. By definition, L_{21} involves both L_2-norm and L_1-norm: L_2-norm computes groups of weights that are associated with features centered at a landmark; there are totally 49 landmarks and thus 49 groups. Meanwhile, L_1-norm performs a sparse selection among 49 groups. As illustrated in Figs. 2 and 8, the selected patches are tentatively sparse using this regularizer.

1) Patch Importance: To validate the ability of maintaining the *specificity* of patches (*i.e.*, how meaningful the selected patches correspond to a particular AU), we compare standard feature learning[2] (treat each feature independently) with our patch learning (treat features as groups). We define the importance of the p-th patch for the ℓ-th AU as its contribution to the classification task, *i.e.*, $\|\mathbf{w}_\ell^p\|_2$. As shown in Fig. 2, compared to feature learning, patch learning offers a better interpretation of important patches in 3 AUs. For instance, patches around inner eyebrow contain higher importance for AU 1; for AU 24, patches around the mouth (especially upper lips) are shown more important. Moreover, compared to previous work that

[1] Bold capital letters denote a matrix $\mathbf{X}$; bold lower-case letters denote a column vector $\mathbf{x}$. $\mathbf{x}_i$ the i-th column of the matrix $\mathbf{X}$. All non-bold letters represent scalars. X_{ij} denotes the scalar in the (i, j)-th entry of the matrix $\mathbf{X}$. x_j denotes the scalar in the jth element of $\mathbf{x}$. $\mathbf{1}_m \in \mathbb{R}^m$ is a vector of ones. $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ are matrices of zeros. $I(x)$ is an indicator function that returns 1 if the statement x is true, and 0 otherwise.

[2] ℓ_1-regularized linear SVM [12] was used as feature learning.

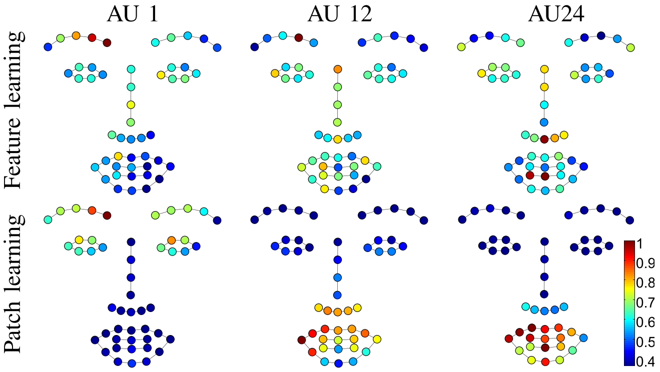


Fig. 2. Patch importance between standard feature learning and our patch learning for AUs 1, 12 and 24 on CK+ dataset. Weights on each patch are computed as the norm of their classification vectors normalized to [0,1].

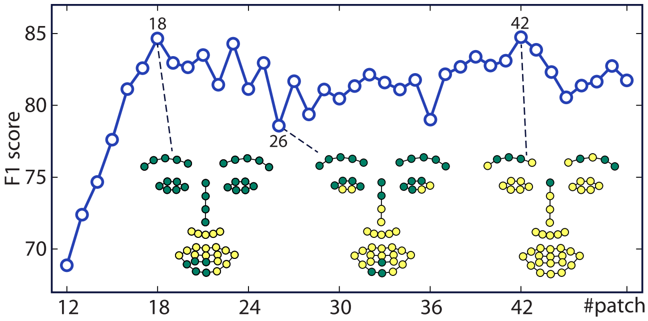
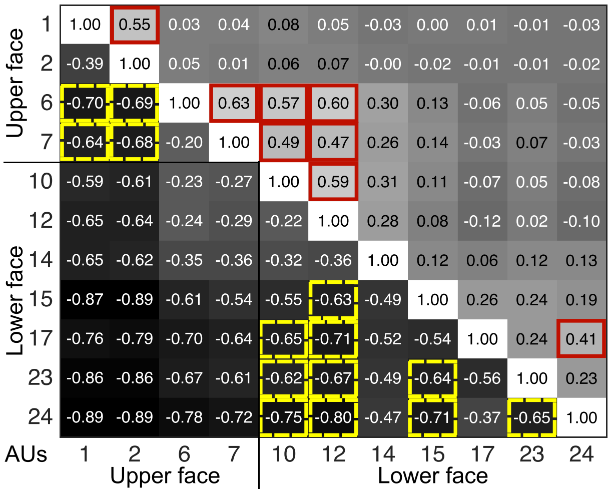


Fig. 3. F1-Norm with respect to different #patches for AU 12 on CK+ dataset. Three marked faces indicate the 18, 26 and 42 selected patches, which are depicted as light yellow circles.

manually defines a fixed region for AU 12 (*e.g.*, [30], [35]), our patch learning for AU 12 automatically emphasizes upper lips (not lower lips), the patches around lower nose, and minor importance on the lower eyelid (corresponding to AU 6). It can be observed that patch learning brings better *specificity* of relevant facial patches.

2) The Number of Patches Versus Performance: A natural question to ask is how the number of patches influences the performance of AU detection. Intuitively, introducing more patches should improve detection performance because more information is provided. To answer this question, we performed an experiment on AU 12 in CK+ dataset. Patches were selected in a descending order with respect to the patch importance. As shown in Fig. 3, the performance increases quickly until it hits the best performance with 18 patches, which associate with the zygomatic major in AU 12 (upper lips and lower nose). When #patches becomes 26, patches on lower eyelid (associated with AU 6) are included, showing that patches associated with AU 6 are related to AU 12. However, the performance drops slightly because not all patches carry useful information for a particular AU, coinciding with the findings [43]. Introducing more patches potentially include more noises that fluctuate the performance. Observing similar performance between #patches=18 and #patches=42, one can justify the importance of patch specificity, *i.e.*, only a subset of patches are discriminative for AU detection.


Upper face
Lower face
Upper face
Lower face

AUs	1	2	6	7	10	12	14	15	17	23	24
1	1.00	0.55	0.03	0.04	0.08	0.05	-0.03	0.00	0.01	-0.01	-0.03
2	-0.39	1.00	0.05	0.01	0.06	0.07	-0.00	-0.02	-0.01	-0.01	-0.02
6	-0.70	-0.69	1.00	0.63	0.57	0.60	0.30	0.13	-0.06	0.05	-0.05
7	-0.64	-0.68	-0.20	1.00	0.49	0.47	0.26	0.14	-0.03	0.07	-0.03
10	-0.59	-0.61	-0.23	-0.27	1.00	0.59	0.31	0.11	-0.07	0.05	-0.08
12	-0.65	-0.64	-0.24	-0.29	-0.22	1.00	0.28	0.08	-0.12	0.02	-0.10
14	-0.65	-0.62	-0.35	-0.36	-0.32	-0.36	1.00	0.12	0.06	0.12	0.13
15	-0.87	-0.89	-0.61	-0.54	-0.55	-0.63	-0.49	1.00	0.26	0.24	0.19
17	-0.76	-0.79	-0.70	-0.64	-0.65	-0.71	-0.52	-0.54	1.00	0.24	0.41
23	-0.86	-0.86	-0.67	-0.61	-0.62	-0.67	-0.49	-0.64	-0.56	1.00	0.23
24	-0.89	-0.89	-0.78	-0.72	-0.75	-0.80	-0.47	-0.71	-0.37	-0.65	1.00

Fig. 4. The AU relation matrix studied on more than 350,000 valid frames with AU labels. Red solid and dashed yellow rectangles, indicate the *positive correlations* and *negative competitions* studied in this work, respectively.

C. Multi-Label Learning

The next task of JPML is to exploit label relations for AU detection. Learning with related labels effectively increases the sample size for each class, and improves the prediction performance (*e.g.*, [3], [38]). Despite of the AU relations derived from prior knowledge [19], [34], this section explores statistically the AU co-occurrence among more than 350,000 annotated video frames. Below we describe how we discover these relations, and how they can be incorporated into JPML.

1) Discover AU Relations: We seek AU relations by statistically analyzing three datasets, CK+ [22], GFT [27] and BP4D [40], which in total contain 214 subjects and more than 350,000 valid FACS-coded frames. The top frequently occurring AUs are used throughout this paper. Here, our goal is to discover likely and rarely co-occurring AUs.

Fig. 4 shows the relation matrix studied on the three datasets. The (i, j)-th entry of the upper right matrix was computed as the coefficient correlation between the i-th and the j-th AU using ground truth frame-level labels; an entry of the lower left matrix was computed on the labels containing at least either the i-th or the j-th AU. One could interpret the upper matrix in Fig. 4 as a mutual relation of concurring AU pairs, and the lower matrix as an exclusive relation that one AU competes against another. After investigating this matrix with the FACS [11] and related studies [19], [34], we derive two types of AU relations, *positive correlation* and *negative competition*, as summarized in Table I.

To discover these relations, we derive explicit rules as follows. AUs with over moderate positive correlations, *i.e.*, correlation coefficient $\geq$ 0.40, are assigned as *positive correlations*, *e.g.*, AUs (6, 12) co-occur frequently to describe a Duchenne smile. AUs with large negative correlations, *i.e.*, the negative correlation coefficient ≤ -0.60, are selected as *negative competitions*, implying these AUs compete against each other and thus avoid occurring at the same time, *e.g.*, AUs (12, 15) have negative influences on each

TABLE I

AU RELATIONS DISCOVERED AND USED IN THIS STUDY

AU relations	**AU groups**
Positive correlation	(1,2), (6,7), (6,10), (7,10), (6,12), (7,12), (10,12), (17,24)
Negative competition	(1,6), (1,7), (2,6), (2,7), (10,17), (10,23), (10,24), (12,15), (12,17), (12,23), (12,24), (15,23), (15,24), (23,24)

*Parentheses indicate pairwise AUs with a particular relation.

other (similar findings in [19]). Note that, for the lower matrix, we exclude the consideration of relations between upper face and lower face AUs, because their facial muscles function separately, and thus do not literally compete against each other. In addition, one can observe that the absolute values of lower matrix are higher than the upper ones, providing another evidence that most AU combinations rarely co-occur [30].

2) Incorporating AU Relations With JPML: Denote the set of AU pairs with positive correlations and with negative competitions as $\mathcal{P}$ and $\mathcal{N}$, respectively. For instance, AUs (1,2) and AUs (6,12) are in $\mathcal{P}$; AUs (15,23), AUs (15,24), and AUs (23,24) are in $\mathcal{N}$. To incorporate the AU relations discovered above, we introduce the *relational regularizer*:

$$\Psi(\mathbf{W},\mathbf{X}) = \beta_1 PC(\mathbf{W},\mathbf{X},\mathcal{P}) + \beta_2 NC(\mathbf{W},\mathbf{X},\mathcal{N}), \qquad (3)$$

where β_1 and β_2 are tradeoff coefficients. $PC(\mathbf{W},\mathbf{X},\mathcal{P})$ captures the AU relations of positive correlations:

$$PC(\mathbf{W},\mathbf{X},\mathcal{P}) = \frac{1}{2}\sum_{(i,j)\in\mathcal{P}} \gamma_{ij}\|\mathbf{w}_i^\top\mathbf{X} - \mathbf{w}_j^\top\mathbf{X}\|_2^2, \qquad (4)$$

where γ_{ij} is a pre-defined similarity score that determines how similar two predictions $\mathbf{w}_i^\top\mathbf{X}$ and $\mathbf{w}_j^\top\mathbf{X}$ are. The larger γ_{ij} is, the more similar predictions are for the i-th and the j-th AUs in $\mathcal{P}$ ($\gamma_{ij} = 2000$ in our experiments). The intuition behind this regularizer is that AUs within the same positive group would share similar predictions, and thus minimizes (4). On the other hand, the regularizer of negative competition $NC(\mathbf{W},\mathbf{X},\mathcal{N})$ is defined in analogy to exclusive lasso [44]:

$$NC(\mathbf{W},\mathbf{X},\mathcal{N}) = \sum_{i=1}^{N}\sum_{n=1}^{|\mathcal{N}|}\left(\sum_{j\in\mathcal{N}_n}\left|\mathbf{w}_j^\top\mathbf{x}_i\right|\right)^2, \qquad (5)$$

where $\mathcal{N}_n$ indicates the n-th AU pair in $\mathcal{N}$, and $|\mathcal{N}| = 14$ (as shown in Table I). For example, $\mathcal{N}_1$ is the AU pair (1,6) with negative competition. Because the ℓ_1 norm in Eq. (5) tends to achieve a sparse solution, if one classifier predicts AU1 in the group $\mathcal{N}_1$, the AU6 classifier tends to generate small prediction values so that the value of the regularizer is minimized. In this way, "competitions" are introduced among the predictions within the same negative group. As a result, we solve for the multi-label learning task of JPML:

$$\min_{\mathbf{W}}\ L(\mathbf{W},\mathcal{D}) + \Psi(\mathbf{W},\mathbf{X}). \qquad (6)$$

We detail our algorithm to solve JPML in the next section.

Algorithm 1 Patch Learning (PL)

```
Input: Training data $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$, ML matrix $\mathbf{W}_2$, Lagrange
    multiplier of ADMM $\rho$ and $\mathbf{U}$, learning rate $\eta_1$, and penalty
    parameter $\alpha$.
Output: PL matrix $\mathbf{W}_1 \in \mathbb{R}^{D\times L}$ with sparse groups of rows.
for $\ell = 1,\ldots,L$ do
    $\mathbf{w}_{1\ell}^{(0)} = \frac{1}{D}\mathbf{1}_D, \mathbf{v}^{(0)} = \frac{1}{D}\mathbf{1}_D, a^{(0)} = 1, t = 0$; // Initialization
    while not convergence do
        $\mathbf{z}^{(t)} = \mathbf{v}^{(t)} - \eta_1(\nabla L(\mathbf{w}_{1\ell}^{(t)},\mathcal{D}) + \mathbf{u}_\ell^{(t)} + \rho(\mathbf{w}_{1\ell}^{(t)} - \mathbf{w}_{2\ell}^{(t)}))$;
        for $p = 1,\ldots,49$ do
            $\mathbf{w}_{1\ell}^{p(t+1)} = I(\|\mathbf{z}^{p(t)}\|_2 > \alpha)(1 - \frac{\alpha}{\|\mathbf{z}^{p(t)}\|_2})\mathbf{z}^{p(t)}$;
            // $\mathbf{w}_{1\ell}^p$ is the $p$-th patch within the $\ell$-th column of $\mathbf{W}_1$
        end for
        $a^{(t+1)} = \frac{2}{t+1}$;
        $\mathbf{v}^{(t+1)} = \mathbf{w}_{1\ell}^{(t+1)} + (\frac{1-a^{(t)}}{a^{(t)}}a^{(t+1)})(\mathbf{w}_{1\ell}^{(t+1)} - \mathbf{w}_{1\ell}^{(t)})$;
        $t = t+1$;
    end while
end for
```

D. Algorithm

Because the patch regularizer $\Omega(\mathbf{W})$ and the relational regularizer $\Psi(\mathbf{W},\mathbf{X})$ constrain on rows and columns of $\mathbf{W}$, respectively, Problem (1) has no analytical solution. Instead, we rewrite (1) by introducing auxiliary variables $\mathbf{W}_1, \mathbf{W}_2$, and then jointly optimize $\mathbf{W}_1$ and $\mathbf{W}_2$ using ADMM [2]:

$$\begin{aligned}\min_{\mathbf{W}_1,\mathbf{W}_2}\ & L(\mathbf{W}_1,\mathcal{D}) + \alpha\Omega(\mathbf{W}_1) + \Psi(\mathbf{W}_2,\mathbf{X})\\ \text{s.t.}\quad & \mathbf{W}_1 = \mathbf{W}_2. \end{aligned} \qquad (7)$$

Introducing a Lagrangian multiplier $\mathbf{U} \in \mathbb{R}^{\mathbf{D}\times L}$ and a penalty term ρ, the augmented Lagrangian can be written as:

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{W}_1,\mathbf{W}_2,\mathbf{U}) = {} & L(\mathbf{W}_1,\mathcal{D}) + \alpha\Omega(\mathbf{W}_1) + \Psi(\mathbf{W}_2,\mathbf{X})\\ & + \langle\mathbf{U},\mathbf{W}_1 - \mathbf{W}_2\rangle + \frac{\rho}{2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F^2,\end{aligned} \qquad (8)$$

which can be then solved with ADMM following the updates:

$$\mathbf{W}_1^{(k+1)} = \min_{\mathbf{W}_1}\ \mathcal{L}_\rho(\mathbf{W}_1,\mathbf{W}_2^{(k)},\mathbf{U}^{(k)}), \qquad (9)$$

$$\mathbf{W}_2^{(k+1)} = \min_{\mathbf{W}_2}\ \mathcal{L}_\rho(\mathbf{W}_1^{(k+1)},\mathbf{W}_2,\mathbf{U}^{(k)}), \qquad (10)$$

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} + \rho(\mathbf{W}_1^{(k+1)} - \mathbf{W}_2^{(k+1)}). \qquad (11)$$

Solving (9) involves the patch regularizer $\Omega(\mathbf{W}_1)$ and the augmented terms in $\mathcal{L}_p$. Because solving for $\mathbf{W}_1$ with $L_{2,1}$ norm is a non-smooth problem, we decompose $L_{2,1}$ norm into 49 sub-problems, which are then solved using the accelerated gradient method [4]. Algo. 1 summarizes the detailed procedure. The convergence condition in the algorithm is $\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_2 \leq \delta$ ($\delta = 10^{-5}$ in our case).

Fig. 5 illustrates the convergence process of PL on AU 12. While the number of iteration increases, PL converges to a subset of patches that preserve better specificity. On iteration #1, many patches irrelevant to AU 12 are selected. From iteration #10 to #30, patches associated with AU 12 are strengthen but still involve unrelated regions such as eyes. PL converges at it#60, revealing the patches around lower nostril

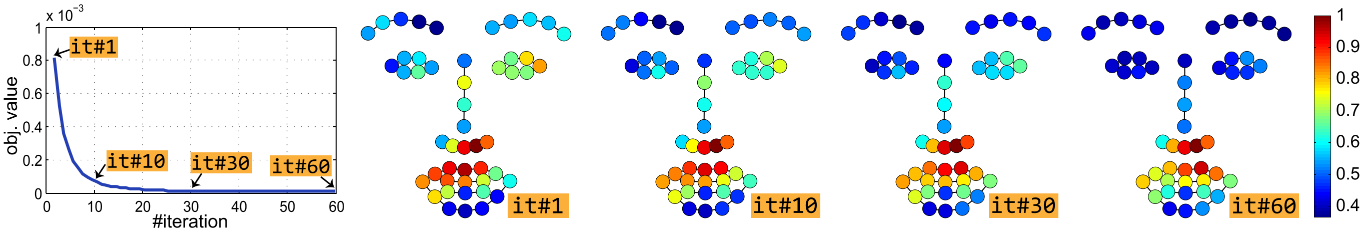


Fig. 5. Illustration of convergence curve on learning active patches on AU 12 with algorithm PL. While the iterations proceed, PL identifies the regions for AU 12 (lip corner puller) with better specificity.

wing and upper mouth, the regions that zygomaticus major muscle triggers for AU 12.

Solving (10) involves the relational regularizer $\Psi(\mathbf{W}_2, \mathbf{X})$ and the augmented terms in $\mathcal{L}_p$. Algo. 2 summarizes the procedure. For $\Psi(\cdot, \cdot)$, the positive correlation $PC(\mathbf{W}_2, \mathbf{X}, \mathcal{P})$ is smooth in $\mathbf{W}_2$, but the negative competition $NC(\mathbf{W}_2, \mathbf{X}, \mathcal{N})$ is not making Problem (10) non-differentiable. Instead of solving (10) directly, we adopt Nesterov's approximation [25] to smooth the objective. Given a training sample $\mathbf{x}_i$ and its negative relation $\mathcal{N}_i$, we denote $\mathbf{W}_{\mathcal{N}_i}$ as a $D \times |\mathcal{N}_i|$ submatrix of $\mathbf{W}$ where each column contains $\mathbf{w}_j$ and $j \in \mathcal{N}_i$. Let $\|\mathbf{W}_{\mathcal{N}_i}^\top \mathbf{x}_i\|_1 = \sum_{j \in \mathcal{N}_i} \left|\mathbf{w}_j^\top \mathbf{x}_i\right|$, we can write its dual norm as $\|\mathbf{w}_j^\top \mathbf{x}_i\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \mathbf{w}_j^\top \mathbf{x}_i, \mathbf{z}\rangle$, and smooth $NC(\mathbf{W}_2, \mathbf{X}, \mathcal{N})$ following the Nesterov's approximation [25]. That is, the $\|\mathbf{w}_j^\top \mathbf{x}_i\|_1$ can be approximated by the following smooth function:

$$q_\mu(\mathbf{w}_j, \mathbf{x}_i) = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \mathbf{w}_j^\top \mathbf{x}_i, \mathbf{z}\rangle - \frac{\mu}{2}\|\mathbf{z}\|_2^2, \tag{12}$$

where μ is a parameter to control the approximation accuracy. For a fixed $\mathbf{w}$, we obtain $\mathbf{z}$ as:

$$\mathbf{z} = \min\left\{1, \max\left\{-1, \frac{\mathbf{w}_j^\top \mathbf{x}_i}{\mu}\right\}\right\}. \tag{13}$$

Based on these preliminaries, we solve the non-smooth objective of problem (10) as the smooth approximation:

$$\begin{aligned} &L(\mathbf{W}_1, \mathcal{D}) + \alpha\Omega(\mathbf{W}_1) + \beta_1 PC(\mathbf{W}_2, \mathbf{X}, \mathcal{P}) \\ &+\beta_2 NC_\mu(\mathbf{W}_2, \mathbf{X}, \mathcal{N}) + \langle \mathbf{U}, \mathbf{W}_1 - \mathbf{W}_2\rangle + \frac{\rho}{2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F^2, \end{aligned} \tag{14}$$

where

$$NC_\mu(\mathbf{W}_2, \mathbf{X}, \mathcal{N}) = \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{N}|} \left(\sum_{j \in \mathcal{N}_n} q_\mu(\mathbf{w}_j, \mathbf{x}_i) \right)^2. \tag{15}$$

JPML is optimized by iterating patch learning (Algo. 1) and multi-label learning (Algo. 2). Because the ADMM form in (8) is bi-convex, *i.e.*, convex in $\mathbf{W}_1$ when $\mathbf{W}_2$ is fixed and vise versa, it is guaranteed to converge to a critical point [13]. Fig. 6 shows the convergence process of JPML. In training, the maximum iteration is set as 30, while JPML typically converges within 5 iterations. As can be seen in Fig. 6(a), for each iteration of PL and ML, JPML manages to keep the averaged error between $\mathbf{W}_1^{(t)}$ and $\mathbf{W}_2^{(t)}$ as low as 10^{-5}. By adding *positive correlations* and *negative competitions* into patch learning, much more accurate correlations closed to ground truth can be learned. Quantitatively, the distance between predictions and ground truth decreased 3.4 times, as shown in Fig. 6(d) and (e). Note that the entries of AUs (1,2) in Fig. 6(c) and (e) are empty because in CK+ AUs (1,2) always co-occur, leading to a zero variance during the computation of correlation coefficient.

Algorithm 2 Multi-Label Learning (ML)

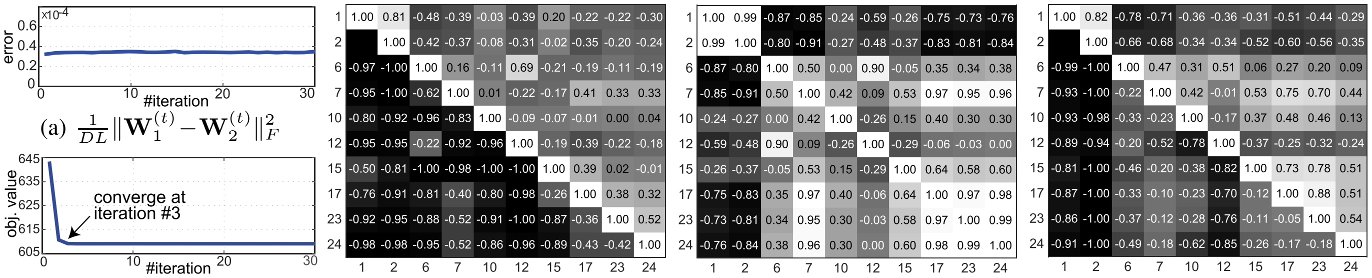
```
Input: Training data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, PL matrix $\mathbf{W}_1$, Lagrange
    multiplier of ADMM $\rho$ and $\mathbf{U}$, learning rate $\eta_2$, penalty parameter $\beta_1, \beta_2$,
    and accuracy control parameter $\mu$.
Output: ML matrix $\mathbf{W}_2 \in \mathbb{R}^{D \times L}$.
$\mathbf{W}_2^{(0)} = \frac{1}{D}\mathbf{1}_{D\times L}, \mathbf{V}^{(0)} = \frac{1}{D}\mathbf{1}_{D\times L}, a^{(0)} = 1, t = 0$; // Init.
while not convergence do
    $\mathbf{U}^{(t)} = (1 - a^{(t)})\mathbf{W}_2^{(t)} + a^{(t)}\mathbf{V}^{(t)}$;
    $\mathbf{H}_\mu = \mathbf{0}_{L\times D}$;
    for $i = 1, \ldots, N$ do
        $\mathbf{z}_i = \min(1, \max(-1, \frac{\mathbf{U}^{(t)\top}\mathbf{x}_i}{\mu}))$;
        $q_i = \mathbf{z}_i^\top \mathbf{U}^{(t)\top}\mathbf{x}_i - \frac{\mu}{2}\|\mathbf{z}_i\|_2^2$;
        $\mathbf{H}_\mu = q_i(\mathbf{z}_i \mathbf{x}_i^\top)$;
    end for
    $\mathbf{V}^{(t+1)} = \mathbf{V}^{(t)} - \frac{1}{\eta_2}(\beta_2\mathbf{H}_\mu^\top + \beta_1\nabla PC(\mathbf{U}^{(t)}) + \rho(\mathbf{W}_1 - \mathbf{U}^{(t)}) - \mathbf{U})$;
    $\mathbf{W}_2^{(t+1)} = (1 - a^{(t)})\mathbf{W}_2^{(t)} + a^{(t)}\mathbf{V}^{(t+1)}$;
    $a^{(t+1)} = \frac{2}{t+1}$;
    $t = t + 1$;
end while
```

IV. Disentangle Common and Specific Patches for Holistic Expression Recognition

Recent studies [21], [43] suggest that only few facial parts contribute to expression disclosure and the evocation of different expressions may involve the muscular movements from the same facial regions. Facial features, therefore, can be disentangled into non-overlapped groups of patches, or the so-called *common* and *specific* patches. As indicated in [21] and [43], these patches usually provide better interpretation and generalization. This section shows that our JPML framework can be generalized for obtaining such patches with comparable performance. Note that different from standard uniformly distributed patches, *e.g.*, [21], [43], we used our patch definition centered at landmarks as discussed in Section III-B. We term this special case of JPML the Common-Specific Active Patch Learning (CS-APL).

A. Learning Common Patches

Common patches are defined as the most informative patches across all expressions [21], [43]. Informed by this



(a) $\frac{1}{DL}\|\mathbf{W}_1^{(t)} - \mathbf{W}_2^{(t)}\|_F^2$

(b) Obj. value of Eq. (8)

	1	2	6	7	10	12	15	17	23	24
1	1.00	0.81	-0.48	-0.39	-0.03	-0.39	0.20	-0.22	-0.22	-0.30
2		1.00	-0.42	-0.37	-0.08	-0.31	-0.02	-0.35	-0.20	-0.24
6	-0.97	-1.00	1.00	0.16	-0.11	0.69	-0.21	-0.19	-0.11	-0.19
7	-0.95	-1.00	-0.62	1.00	0.01	-0.22	-0.17	0.41	0.33	0.33
10	-0.80	-0.92	-0.96	-0.83	1.00	-0.09	-0.07	-0.01	0.00	0.04
12	-0.95	-0.95	-0.22	-0.92	-0.96	1.00	-0.19	-0.39	-0.22	-0.18
15	-0.50	-0.81	-1.00	-0.98	-1.00	-1.00	1.00	0.39	0.02	-0.01
17	-0.76	-0.91	-0.81	-0.40	-0.80	-0.98	-0.26	1.00	0.38	0.32
23	-0.92	-0.95	-0.88	-0.52	-0.91	-1.00	-0.87	-0.36	1.00	0.52
24	-0.98	-0.98	-0.95	-0.52	-0.86	-0.96	-0.89	-0.43	-0.42	1.00

(c) Relation matrix in CK+

	1	2	6	7	10	12	15	17	23	24
1	1.00	0.99	-0.87	-0.85	-0.24	-0.59	-0.26	-0.75	-0.73	-0.76
2	0.99	1.00	-0.80	-0.91	-0.27	-0.48	-0.37	-0.83	-0.81	-0.84
6	-0.87	-0.80	1.00	0.50	0.00	0.90	-0.05	0.35	0.34	0.38
7	-0.85	-0.91	0.50	1.00	0.42	0.09	0.53	0.97	0.95	0.96
10	-0.24	-0.27	0.00	0.42	1.00	-0.26	0.15	0.40	0.30	0.30
12	-0.59	-0.48	0.90	0.09	-0.26	1.00	-0.29	-0.06	-0.03	0.00
15	-0.26	-0.37	-0.05	0.53	0.15	-0.29	1.00	0.64	0.58	0.60
17	-0.75	-0.83	0.35	0.97	0.40	-0.06	0.64	1.00	0.97	0.98
23	-0.73	-0.81	0.34	0.95	0.30	-0.03	0.58	0.97	1.00	0.99
24	-0.76	-0.84	0.38	0.96	0.30	0.00	0.60	0.98	0.99	1.00

(d) Patch learning

	1	2	6	7	10	12	15	17	23	24
1	1.00	0.82	-0.78	-0.71	-0.36	-0.36	-0.31	-0.51	-0.44	-0.29
2		1.00	-0.66	-0.68	-0.34	-0.34	-0.52	-0.60	-0.56	-0.35
6	-0.99	-1.00	1.00	0.47	0.31	0.51	0.06	0.27	0.20	0.09
7	-0.93	-1.00	-0.22	1.00	0.42	-0.01	0.53	0.75	0.70	0.44
10	-0.93	-0.98	-0.33	-0.23	1.00	-0.17	0.37	0.48	0.46	0.13
12	-0.89	-0.94	-0.20	-0.52	-0.78	1.00	-0.37	-0.25	-0.32	-0.24
15	-0.81	-1.00	-0.46	-0.20	-0.38	-0.82	1.00	0.73	0.78	0.51
17	-0.87	-1.00	-0.33	-0.10	-0.23	-0.70	-0.12	1.00	0.88	0.51
23	-0.86	-1.00	-0.37	-0.12	-0.28	-0.76	-0.11	-0.05	1.00	0.54
24	-0.91	-1.00	-0.49	-0.18	-0.62	-0.85	-0.26	-0.17	-0.18	1.00

(e) JPML

Fig. 6. Illustration of JPML on the CK+ dataset: (a) Error of $\frac{1}{DL}\|\mathbf{W}_1^{(t)} - \mathbf{W}_2^{(t)}\|_F^2$ v.s. #iteration (t), (b) objective value in (8) v.s. #iteration, (c) ground truth relation matrix (correlation coefficients between ground truth AU labels), (d) relation matrix at the initialization step (with patch learning only), and (e) relation matrix computed by predictions of JPML. The difference of correlation coefficient between (c) and (d) is 0.51, and that between (c) and (e) is 0.15, showing that JPML helps preserve the relations between AUs.

definition, we hypothesize that different expressions may share a subset of facial patches. Recall that patch learning in Section III-B is a single-task learning for individual AUs. To model common patches, we reformulate the single-task patch learning as a multi-task learning problem by rewriting the patch regularizer:

$$\Omega(\mathbf{W}) = \sum_{p=1}^{49} \left\|[\mathbf{w}_1^p, \ldots, \mathbf{w}_L^p]\right\|_2, \tag{16}$$

where $\mathbf{W} \in \mathbb{R}^{D \times L}$ is the classification matrix for predicting L expressions, and $\mathbf{w}_\ell^p$ is the classification vector with respect to the p-th patch and the ℓ-th expression. This regularizer selects patches across L expressions as one group, leading to totally 49 groups. Due to the group-sparsity, the elements in the same group can or cannot be selected simultaneously. This regularizer, thus, encourages a sparse selection of patches that are commonly important for all expressions. Common patches are selected as the groups with the largest patch importance $\|[\mathbf{w}_1^p, \ldots, \mathbf{w}_L^p]\|_2$. Note that different from Sec. III-B, here the patch importance is defined across L expressions ($L = 6$).

B. Learning Specific Patches

In contrast to common patches, *Specific patches* are defined as the patches contributing to only one particular expression. Different from learning common patches over all 6 expressions, specific patches of facial expressions are learned in pairs ($L = 2$). Learning specific patches using the JPML framework can be characterized into the following steps. First, we solve patch learning with the regularizer in (16), and select the common patches as the top ones ranked by patch importance. Then, similar to [43], excluding the overlapping patches, the remaining ones are considered as the specific patches. More discussion and results will be shown in Section V-E.

V. Experiments

This section evaluates the proposed JPML framework in several scenarios: Within-dataset AU detection, cross-dataset AU detection, qualitative results of patch learning, and holistic expression recognition. We commence by experimental settings and close by discussion.

A. Settings

1) Datasets: We evaluated JPML in three datasets that involve posed and spontaneous facial behaviors in varied contexts. Each database has been FACS-coded by well-experienced coders. Inter-observer agreement in each was quantified using coefficient κ, which controls for chance agreement between coders, and it was maintained at $\kappa \geq 0.8$, which represents high inter-observer agreement. Because severely skewed base rates attenuate estimates of classifier performance [15], only AUs occurring more than 3% to 5% of the time were included for analysis. Across datasets, 10 to 11 AUs met this criterion.

(1) CK+ [22] is a leading testbed for facial expression analysis. It consists of 593 sequences of posed facial actions from 123 subjects. The last frames of each sequence were selected as positive samples. In all, 593 images with 10 AUs were used.

(2) GFT [27] consists of social interaction between 720 previously unacquainted young adults that were assembled into groups of three persons each and observed over the course of a 30-minute group formation task. Moderate out-of-plane head motion and occlusion are presented in the videos which makes the AU detection challenging. Videos from 50 participants with AU coding were used. For each AU, we randomly sampled 100 positive frames and 200 negative frames from each subject for the training set.

(3) BP4D [40] contains spontaneous facial expressions in 41 young adults during various emotion inductions while interacting with an experimenter. Three hundred twenty-eight videos (41 participants×8 videos each) were used. For each AU, we randomly sampled 50 positive frames and 100 negative frames from each subject for the training set.

Even though AUs with very low base rates were omitted, skewness nevertheless varied considerably. To control for the effects of skewness on AU detection, test statistics were normalized using the procedure of [15]. By normalizing for skewness we were able to reliably compare results within and

TABLE II

SKEWNESS ON EACH AU WITHIN DIFFERENT DATASETS

AU	1	2	6	7	10	12	14	15	17	23	24
CK+	1.5	2.3	2.4	3.1	20.8	3.1	9.2	8.9	1.8	6.6	6.6
GFT	10.1	8.3	2.1	1.5	1.5	2.0	0.6	9.7	2.7	4.9	8.4
BP4D	3.8	4.9	1.2	0.8	0.7	0.8	1.1	4.9	1.9	5.0	5.5

TABLE III

COMPARISONS ON THE CK+ DATASET. BRACKETED NUMBERS STAND FOR THE BEST PERFORMANCE; BOLD NUMBERS FOR THE SECOND BEST.

	F1-Norm							
AU	SP-SVM	SPSD	LSVM	LL1	MT-MKL	ML	APL	JPML
1	61.8	44.4	85.8	83.4	73.0	**89.0**	86.4	[**90.0**]
2	63.9	47.9	90.9	87.4	87.8	**92.7**	86.6	[**93.0**]
6	61.7	34.2	**75.3**	[**76.2**]	61.9	70.7	70.5	74.2
7	60.0	50.8	[**70.8**]	**70.0**	–	61.6	62.8	66.7
12	65.5	47.4	[**80.7**]	80.0	73.3	75.2	76.6	[**80.7**]
14	66.3	59.9	**67.7**	**67.7**	–	–	[**69.5**]	–
15	65.8	53.4	67.5	66.7	67.8	61.9	**79.2**	**73.6**
17	60.4	62.2	**80.5**	**80.5**	68.3	80.3	80.0	[**83.5**]
23	66.2	65.0	69.3	69.8	–	69.7	**83.5**	**74.3**
24	68.3	65.8	71.1	**71.4**	–	67.5	**75.9**	65.9
AA.	64.0	53.1	**76.0**	75.3	–	–	[**77.1**]	–
AR.	63.7	52.3	76.9	76.2	72.0	74.3	**77.0**	[**78.0**]

between datasets. Table II summarizes the skew factor defined as the ratio of the number of negative samples to the number of positive ones.

2) Pre-Processing: IntraFace [8] was used to track 49 facial landmarks. Tracked landmarks were registered to a reference face using similarity transform. Appearance features were extracted using SIFT descriptor [45] at frame level, resulting in 49×128-D features for each image. To take full advantage of the datasets, we divided GFT and BP4D into 10 splits of independent participants. Because CK+ only contains 593 images, 5 splits were adopted. Each experiment was conducted using a leave-one-split-out protocol, *i.e.*, train on $(n-1)$ splits and test on the remaining one. Results were reported after all n splits were evaluated.

3) Evaluation Metrics: We used two metrics F1-Norm (frame-based) and F1-Event (segment-based). F1-Norm [15] is computed by multiplying false negatives and true negatives by the factor of skewness, which is computed as the ratio of positive samples over negative ones. F1-Norm skew-normalizes the standard F1 metric and enables comparison both within and between datasets. On the other hand, F1-Event [9] serves as a segment-based metric defined as the harmonic mean between event-based recall ER and event-based precision EP: F1-Event $= \frac{2 \cdot ER \cdot EP}{ER+EP}$. For each method, we computed the averaged metric over all AUs (denoted as *AA.*), and averaged over only the AUs with relations (denoted as *AR.*).

4) Alternative Methods: To investigate the benefits of JPML, we compared it with baseline methods that include neither Patch Learning (PL) nor Multi-label Learning (ML), and methods that focus on either PL or ML alone. For baseline methods, we trained a Linear SVM (LSVM) [12] on individual AU. As a baseline for feature learning, we used L1-regularized logistic regression (LL1) [12]. Neither LSVM nor L1-regularized logistic regression used patches.

For PL methods, we chose three for comparison. The first is SP-SVM, a baseline SVM trained on manually-defined patches. These patches are informed by observations or by FACS (*e.g.*, [5], [37], [45]). Specifically, the patches used in SP-SVM were defined as landmarks #1∼#10 for AUs 1, 2 and 7, #11∼#30 for AU 6, #11∼#19 for AUs 11 and 14, #32∼#49 for all lower-face AUs. See patch indices in Fig. 1(c). For state-of-the-art PL methods, we compared Structure Preserving Sparse Decomposition (SPSD) [30] and Multi-task Sparse Learning (MTSL) [43]. For SPSD, we computed K-SVD [24] to learn AU atoms on the manually-defined patches. Unlike the two-layer structure used in the original work, we used one layer of SPSD to learn an AU dictionary, because GFT and BP4D do not contain expressions labels. In our experiments, we implemented MTSL using as our patch learning algorithm (Algo. 1) and patches centered at facial landmarks. Note that the original MTSL was defined on emotion bases using uniform segmentation on face images. Following the naming convention [43], we term this variation as Active Patch Learning (APL).

For ML methods, we compared JPML with MT-MKL [41] using RBF and polynomial kernels with the implementation provided by the authors. Because MT-MKL involves computing multiple kernel matrices, it is computationally prohibitive for use with large datasets such as GFT and BP4D. For this reason, MT-MKL was carried out only on CK+. Following [41], we employed 3 AU groups: AUs (1,2), (6,12), and (15,17). We term our multi-label learning (Algo. 2) as ML. We tuned the parameters in Algos. 1 and 2 as follows: α is cross-validated within $\{10^{-3}, 10^{-4}, 10^{-5}\}$, $\eta_1 = 10^{-4}$, $\gamma = 2000$, $\mu = 10^{-4}$, $\eta_2 = 2000$, $\beta_1 = 10^{-3}$, and $\beta_2 = 10^{-4}$.

B. Within-Dataset Action Unit Detection

This section evaluates JPML and alternative methods for within-dataset AU detection. Tables III∼V present results in CK+, GFT, and BP4D, respectively. 10 AUs were evaluated in CK+ dataset; 11 AUs were evaluated in GFT and BP4D. In total, 32 AUs were used cross all three datasets. AUs *without* relationships are underlined. We excluded these AUs for ML and JPML and denoted their results as "–". Because each AU is pre-segmented in CK+ and only peak frames are AU coded, we quantified performance using only F1-Norm. For GFT and BP4D where all video frames contain AU labels, we reported both F1-Norm and F1-Event. Below we discuss the results from three perspectives: patch learning, multi-label learning, and the proposed joint framework JPML.

1) Patch Learning (PL): This paragraph addresses the question: Does APL improve performance against feature learning and alternative PL methods? Across all three datasets, we evaluated a total of 32 AUs with F1-Norm, and 22 with F1-event.

TABLE IV

COMPARISONS ON THE GFT DATASET. BRACKETED NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1-Norm							F1-Event						
AU	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML
1	29.9	33.0	53.0	52.0	[**66.7**]	44.1	**58.0**	**17.8**	12.2	[**20.6**]	17.8	11.5	11.5	15.9
2	60.2	34.7	51.3	45.1	[**64.4**]	43.6	**63.2**	[**21.2**]	12.9	19.6	16.8	12.5	**16.6**	15.0
6	[**77.2**]	34.8	74.7	75.2	57.3	**77.2**	[**79.6**]	46.6	21.6	33.2	42.3	25.5	**50.3**	[**50.8**]
7	56.5	40.3	**72.7**	70.5	67.6	[**73.6**]	[**73.6**]	41.3	25.3	38.2	34.4	34.3	**47.9**	[**54.7**]
10	74.6	41.8	75.8	**77.5**	–	[**78.6**]	–	**45.6**	30.7	41.2	37.9	–	**50.2**	–
12	77.1	76.2	79.2	80.2	67.1	**81.3**	[**84.1**]	47.9	**48.6**	47.9	48.4	15.3	[**53.6**]	46.7
14	64.1	**68.9**	68.5	[**70.4**]	–	66.7	–	42.1	49.0	42.1	**55.0**	–	[**60.6**]	–
15	47.2	30.1	45.8	65.3	66.3	[**67.1**]	**66.2**	16.4	10.6	**39.1**	[**39.7**]	17.8	18.9	**37.9**
17	51.8	32.8	47.6	46.8	67.1	[**74.5**]	**72.0**	33.8	22.9	38.3	**38.9**	27.1	[**48.7**]	38.8
23	49.7	35.9	38.8	43.5	[**66.9**]	**63.9**	60.0	25.9	18.0	**35.4**	28.4	28.6	35.0	[**37.6**]
24	51.1	35.3	56.6	59.2	67.1	**79.0**	[**79.3**]	18.7	12.9	**27.3**	25.0	26.7	19.2	[**35.5**]
AA.	58.1	42.3	59.8	55.4	–	[**68.1**]	–	32.5	23.4	34.2	34.6	–	[**37.5**]	–
AR.	55.6	39.4	57.0	51.3	65.6	**67.1**	[**70.7**]	30.0	19.7	32.5	32.0	22.1	**33.5**	[**37.0**]

TABLE V

COMPARISONS ON THE BP4D DATASET. BRACKETED NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1-Norm							F1-Event						
AU	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML
1	22.9	27.6	40.6	35.6	[**58.6**]	**56.0**	55.5	9.5	10.6	13.0	11.7	14.9	**17.1**	[**17.5**]
2	15.8	15.8	32.1	24.1	56.9	**60.2**	[**62.7**]	7.6	7.6	11.5	10.3	14.0	**16.0**	[**17.2**]
6	45.5	54.6	59.4	**75.2**	62.9	75.0	[**75.7**]	21.9	27.9	17.2	21.1	15.6	[**32.7**]	**30.0**
7	44.1	56.0	55.7	[**70.5**]	66.7	64.3	**66.7**	22.9	**30.5**	20.5	23.6	17.1	[**33.7**]	26.3
10	50.1	55.6	63.0	[**74.3**]	–	**72.9**	–	29.7	32.8	22.0	**34.1**	–	**41.0**	–
12	46.5	54.9	62.5	82.0	67.1	[**82.3**]	**81.4**	28.4	30.6	23.4	25.0	20.5	[**41.3**]	**31.6**
14	44.2	52.7	51.5	61.2	–	**66.0**	–	19.3	28.3	23.5	29.3	–	**29.8**	–
15	13.2	40.5	49.6	56.3	66.0	[**68.4**]	**65.9**	**23.4**	22.9	23.9	18.6	20.4	13.1	[**30.1**]
17	42.3	46.9	40.3	63.4	66.7	[**69.2**]	**65.3**	19.3	21.9	21.2	25.6	20.8	[**33.5**]	**29.4**
23	11.3	23.9	42.1	57.2	67.1	[**68.0**]	**65.2**	**19.4**	19.6	[**21.8**]	19.0	20.6	16.2	[**27.7**]
24	7.3	47.3	21.3	69.5	66.7	[**78.1**]	**77.3**	17.7	18.4	19.0	[**23.1**]	20.4	13.2	[**26.4**]
AA.	31.2	43.2	47.1	59.5	–	[**69.1**]	–	19.9	**22.8**	19.7	23.1	–	[**26.1**]	–
AR.	27.6	40.8	44.8	57.7	64.3	[**69.0**]	**68.4**	18.9	21.1	19.0	21.1	18.3	**24.1**	[**26.2**]

APL outperformed feature learning (LL1 and LSVM) in 26 of 32 with F1-Norm, and 14 of 22 with F1-event. Compared to alternative PL approaches (SP-SVM and SPSD) that used uniformly distributed patches, APL performed better in 30 of 32 comparisons with F1-Norm, and 17 of 22 with F1-event. A possible explanation is that APL defined patches around facial landmarks, and thus better adapted to spontaneous expressions that involve motions, such as lower-face AUs. As can be observed, compared to uniform patches used in alternative methods, APL performed more effectively on lower-face AUs.

2) Multi-Label Learning: This paragraph discusses the benefits of multi-label learning, which models explicitly the relations between AUs. Closest to our work is MT-MKL that assumes classifiers within the same AU group behave similarly. On the contrary, our ML (Sec. III-C) considers positive correlation as well as negative competition on labels (instead of classifiers), and thus more naturally fits the problem in hand; yet MT-MKL only considers positive correlation. In Table III, averaging F1-Norm over the 6 AUs we implemented for MT-MKL, ML outperformed against MT-MKL by 8.8%. In Tables IV and V, we have seen that ML consistently outperforms standard binary classifiers (LL1, LSVM, SPSD and SP-SVM). The improvement is more obvious for AUs with large skewness, such as AUs 1, 2 and 23 in GFT, showing that relations between AU labels are essential to assist AU detection.

3) JPML: Compared to the baselines, APL or ML alone has shown better performance over three datasets. This paragraph focuses discussion on JPML, the proposed joint framework that considers both PL and ML. In all, JPML achieved the best or second best for 22/27 AUs in F1-Norm and for 12/18 AUs for F1-event. In Table III, JPML performed the best for AUs 1, 2, 12 and 15, and improved about 1.3% and 5.0% than APL and ML respectively for F1-norm. It improved more than 7.3% and 7.8% for F1-Norm, and 13% and 67% for F1-event than APL and ML respectively. In Tables IV and V, as more spontaneous expressions were involved, the improvement became more obvious. In all, JPML achieved the highest overall scores in five comparisons using the three datasets. In BP4D, APL was slightly higher than JPML for some AUs. In no cases, the other methods matched or exceeded APL

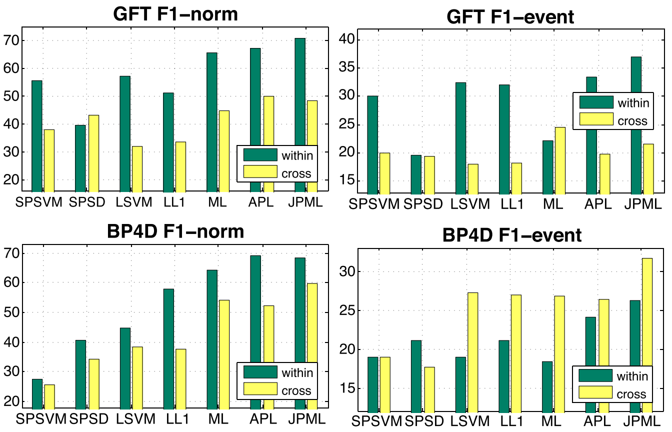


Fig. 7. Comparisons between within-dataset and cross-dataset across all methods. "cross" in GFT means BP4D→GFT, for BP4D it is GFT→BP4D.

and JPML. This suggests our patch-based approach is able to model discriminative information for AU detection, and further boost the performance with an additional ML. In addition, JPML yielded improvements in AUs with larger skew (*e.g.*, AU 1 and AU 2 in GFT and BP4D), as shown in Table II.

C. Cross-Dataset Action Unit Detection

To evaluate the generalizability of the proposed JPML against alternative methods, this section carries out two experiments using the *cross-dataset* scenario. In particular, we used GFT and BP4D by alternately training classifiers on one dataset and test on the other, denoted as "GFT→BP4D" and "BP4D→GFT", respectively. Pro-processing, metrics, and parameter settings follow Sec. V-A except that we trained with all available data from one dataset instead of previously 10 splits. Tables VI and VII show the results.

Observing the performance of APL, we found its F1-norm is about 30% higher than baseline methods for both GFT→BP4D and BP4D→GFT. In specific, on both experiments, for F1-norm, APL outperformed standard feature selection methods (LL1 and LSVM) in 17/22 AUs, and manually defined patch-based methods (SPSVM and SPSD) in 17/22 AUs. For F1-event, APL outperformed SPSVM and SPSD in 15/22 AUs. Comparing the cross-dataset experiments against within-dataset ones (Sec. V-B), the performance of AUs 1 and 2 decreased dramatically for LSVM and LL1, but not for SPSVM, SPSD and APL. This suggests that patch-based methods generalized better than standard feature learning methods. As Fig. 8 shows, the patches automatically selected by APL are visually similar to hand-selected patches in SPSVM and SPSD. The automatically selected patches are close to human's selection, and preserve different importance for each patch. All serves as an evidence that the patch selection ability of APL yields better generalizability than standard feature learning or methods based on manually defined patches, coinciding with the findings in Sec. III-B1 and [43].

Given 9 AUs with known relations, ML generally outperformed baseline methods that took no AU relations into consideration, such as feature learning and algorithms based on manually defined patches. For AUs with large skewness, *i.e.*, AUs 1, 2, 15, 17, 23 and 24, ML improved baselines methods more than 50% in F1-norm and 30% in F1-event. The results justify that AU relations employed by ML help improve performance especially for imbalanced AUs. Specifically, comparing to baseline methods, in F1-norm ML outperformed SPSVM and SPSD in 12/18 AUs, and LSVM and LL1 in 16/18 AUs. For F1-event, ML outperformed SPSVM and SPSD in 16/18 AUs. This validates our statistically derived AU relations can generalize from one dataset to another.

JPML jointly models ML and APL, and from the two cross-dataset experiments, consistently improves overall performance against alternative methods. As can be observed from Tables VI and VII, in F1-norm, taking the positively correlated AUs 6 and 7 for example, APL outperformed others in BP4D→GFT but not in GFT→BP4D; JPML was able to achieve comparable performance on both experiments. In addition, for the negatively competitive AUs 23 and 24, ML outperformed APL in BP4D→GFT, while JPML consistently retains the best performance. Overall, JPML achieved the best and the second best F1-norm and F1-event in 16/18 AUs and in 15/22 AUs in GFT→BP4D and BP4D→GFT, respectively. Considering both the within-dataset and cross-dataset experiments, we observed that JPML outperformed ML or APL alone for AU detection.

Comparing the previous *within-dataset* scenario, as illustrated in Fig. 7, the cross-dataset scenario exhibits overall lower results due to the dataset shift. Interestingly, among the cross-dataset experiment, we found GFT→BP4D generally yields higher F1-event than within-dataset scenario. One possible explanation is because GFT has about 18% more training subjects than BP4D (50 vs 41 subjects), leading to a broader diversity in the training set. Another explanation is that in BP4D many videos contain only negative labels for some AUs. In this sense, predicting temporally consistent events are relatively easy, and thus F1-event tends to be slightly higher than F1-norm. Finally, because F1-event captures the agreement between the ground truth events and the predicted events, we found in our results that F1-event prefers a frequent occurrence of positive predictions (in this case the event-based recall is high), which instead leads to a low F1-norm due to the low frame-based recall.

D. Qualitative Results of Patch Learning

To provide a deeper understanding of the JPML method, this section visualizes the learned patches in terms of AU detection and holitic expression recognition.

1) Learning Patches for AU Detection: Fig. 8 shows the qualitative results of our APL approach. Below we summarize our reasonings for each AU.

- AUs 1 and 2: For both AUs, APL identified important patches around eyes and brows. Inner brow for AU 1 was identified, emphasizing the appearance changes by pulling inner eyebrows. For AU 2, APL learned more importance on upper eyelids, suggesting an eye cover fold

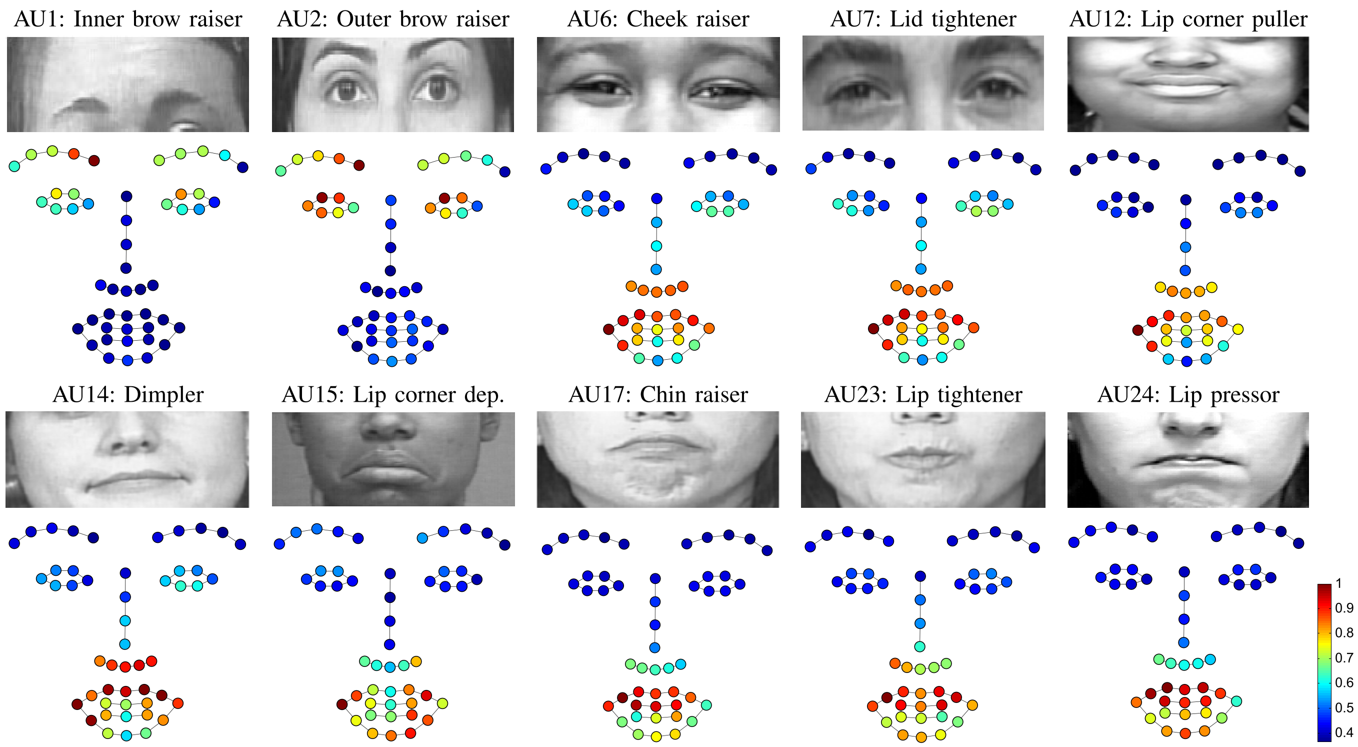


Fig. 8. Visualization of patch importance for AUs generated from the GFT dataset [27]. Weights on each patch are computed as the norm of their classification vectors, and then normalized to [0,1]. The warmer the patch represents, the more importance it is.

TABLE VI
COMPARISONS ON BP4D→GFT. BRACKETED NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1-Norm							F1-Event						
AU	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML
1	26.7	66.7	2.6	0.6	34.0	[**45.4**]	**44.7**	12.1	11.5	4.0	2.0	11.0	[**24.1**]	**15.5**
2	41.4	42.5	8.1	10.6	34.0	**42.5**	[**49.1**]	[**15.3**]	13.6	5.3	4.8	11.0	**15.1**	14.8
6	26.9	56.3	46.3	53.7	63.5	**65.4**	[**71.2**]	18.8	18.9	16.4	17.1	[**29.5**]	16.1	**20.1**
7	54.9	64.8	43.8	50.7	66.0	[**71.9**]	**71.2**	[**26.7**]	21.6	19.5	20.6	**24.3**	17.4	22.9
10	49.7	**53.6**	35.9	42.6	–	[**79.0**]	–	[**35.9**]	**26.6**	20.5	21.7	–	18.6	
12	14.5	25.1	65.5	76.3	65.7	[**73.5**]	**71.4**	26.7	25.7	24.9	26.1	[**36.7**]	19.6	**27.5**
14	10.9	19.1	39.8	**41.4**	–	[**45.6**]	–	11.0	**24.7**	24.0	[**25.0**]	–	20.3	–
15	18.0	7.9	20.5	14.7	**37.1**	[**59.6**]	36.5	22.0	23.0	22.3	23.0	**23.4**	19.9	[**25.1**]
17	34.8	48.1	**55.6**	[**57.0**]	36.2	33.2	32.1	20.9	21.4	**24.1**	[**24.7**]	24.4	22.6	24.0
23	[**59.4**]	34.0	22.0	18.6	**36.1**	31.0	31.4	20.7	20.4	22.9	[**23.4**]	22.1	22.3	**22.7**
24	[**61.9**]	**41.9**	23.4	18.6	28.8	26.1	28.6	22.5	18.2	22.2	**22.7**	[**25.5**]	21.9	21.9
AA.	36.5	**41.8**	33.0	35.0	–	[**52.1**]	–	[**20.9**]	**20.5**	18.7	19.2	–	19.8	–
AR.	37.7	43.0	32.0	33.4	44.6	[**49.8**]	**48.5**	20.4	19.4	18.0	18.3	[**24.6**]	19.9	**21.6**

stretching in a strong AU2. Note that the lateral part of the brow could be pulled by AU 2, and thus the inner corner of brows could move a bit, as suggested by the discovered patches.

- AU 12: APL emphasized upper lip and lower nose, the regions covering deepened nasolabial furrow. We also observed that patches selected for AUs 6 and 7 are similar to 12, suggesting AU 12 is an important source that triggers, or co-occurs, frequently with AUs 6 and 7. This observation coincides with the AU relations discovered in Table I.
- AU 14: AU 14 causes appearance changes of flatten lips and lip corner wrinkling, as suggested by the important patches on upper lips. Interestingly, APL also emphasized the patches on lower nose, capturing the appearance change of stretching nostril wing while 14 was present.
- AUs 15, 23 and 24: These AUs describe the depressor, tightener, and pressor of lips. In contrast to AU 14, in these three scenarios, APL identified important patches around mouth, but not lower nose.
- AU 17: Informed by FACS, the criteria for AU 17 in 17+23 are the same as for 17 alone. Here, APL identified similar patches for AU 17 as for AU 23. An explanation for the similar patches distribution is that some of accompanying signs of the lower lip pushed up by AU 17 are altered by the presence of AU 23.

TABLE VII

COMPARISONS ON GFT→BP4D. BRACKETED NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1-Norm							F1-Event						
AU	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML	SPSVM	SPSD	LSVM	LL1	ML	APL	JPML
1	13.3	[**42.2**]	7.8	6.5	22.3	28.1	**34.0**	8.8	**14.4**	5.8	4.5	11.4	11.9	[**15.1**]
2	21.2	19.1	8.0	6.7	22.1	[**51.6**]	**23.2**	10.3	12.4	6.4	5.6	[**14.2**]	**14.0**	13.6
6	17.4	67.3	65.9	65.6	[**74.5**]	67.3	**74.0**	10.0	13.8	25.7	25.7	[**28.6**]	23.7	**26.2**
7	10.8	62.2	61.8	62.0	[**66.7**]	59.5	**69.9**	10.0	22.3	**31.2**	31.4	30.2	28.7	[**32.0**]
10	39.7	52.8	[**63.2**]	**62.9**	–	58.1	–	29.1	22.8	**35.4**	[**35.6**]	–	31.2	–
12	39.4	23.0	75.1	73.6	[**78.6**]	70.2	**75.8**	28.5	23.5	[**38.3**]	38.4	35.5	33.7	**37.4**
14	15.5	7.6	**57.2**	[**57.4**]	–	44.3	–	11.5	22.0	**40.2**	[**40.4**]	–	33.6	–
15	29.0	27.9	22.9	23.5	**66.3**	53.6	[**63.5**]	22.0	19.7	36.7	**36.9**	31.6	32.2	[**39.6**]
17	55.0	38.6	38.4	38.8	[**66.7**]	47.1	**65.2**	29.3	18.8	34.6	**34.7**	31.9	32.3	[**41.0**]
23	9.2	13.5	31.7	29.1	41.7	**42.8**	[**65.7**]	25.6	17.3	**32.8**	32.7	28.7	31.0	[**40.9**]
24	35.7	14.6	34.3	34.5	46.7	**51.3**	[**70.3**]	23.9	16.2	**32.0**	31.9	28.7	30.5	[**38.6**]
AA.	26.0	33.5	**42.4**	41.9	–	[**52.2**]	–	19.2	18.5	[**29.0**]	**28.9**	–	27.5	–
AR.	25.7	34.2	38.4	37.8	**54.0**	52.4	[**59.9**]	19.0	17.6	**27.2**	26.9	26.8	26.4	[**31.6**]

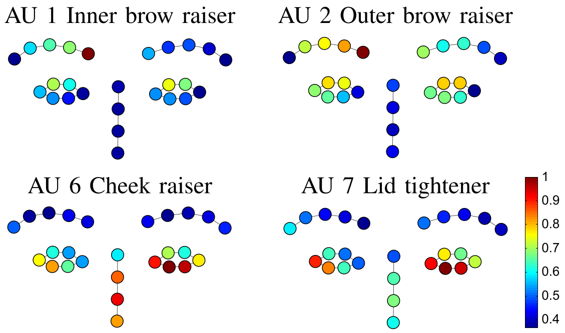


Fig. 9. Visualization of patch importance for upper AUs from the GFT dataset [27] (see illustration of each AU in Fig. 8).

From the visualization in Fig. 8, the patches around lower face contain more information to discriminate lower-face AUs. However, it is unclear for upper-face AUs. To further study the distribution of patches importance on upper faces, we performed a complementary experiment only on upper-face AUs and their associated patches. Fig. 9 illustrates the results. In contrast to Fig. 8, we drew several interesting observations below. We conjecture that large variation of lower face may interfere the identification of patches on upper face.

- AU 6: APL identified patches on lower lids and nose that show the appearance changes of face as FACS describes. However, in Fig. 8, patches around cheek were less emphasized probably due to the interference of lower face AUs.
- AU 7: Fig. 9 clearly tells the importance on lower eyelids, due to appearance change of AU 7 as informed by FACS. Compared to Fig. 8, patches on eyebrows and nose were shown less meaningful.
- AUs 1 and 2: Patch importance distributions are similar between Figs. 8 and 9.

2) Learning Patches for Holistic Expressions: Patch learning also can be used to learn discriminative patches for holistic facial expressions, since they can be described by the combination of AUs as shown in Table VIII. In particular, with the AU labels replaced by expression labels, we applied patch learning algorithm in Algo. (1). The patch importances learned are shown in Fig. 10.

Fig. 10 represents the patch importance distribution learned on holistic expressions using the CK+ dataset. Based on the patch importance, we observe the relations between the patches identified by APL and the AU composition rule as shown in Table. VIII. We discuss these relations below.

- Anger: APL identifies patches on eyebrows, eyes, and outer mouth. From Fig. 8, eyes and outer lips are identified for AU 7 and AU 23 separately. Thus in patches importance level, patches identified for AU 7 and AU 23 are included in important patches for anger.
- Disgust: For a disgust face, APL identifies patches on nose and mouth. The highest patch importance happened on the center of nose, corresponding to the active region of AU 9 (nose wrinkler). In addition, the learned patch importance on the lower nose and lips corners showed a similar pattern with AU 15 in Fig. 8. The correspondence to AUs 9 and 15 complies with the AU composition rule for the disgust expression as shown in Table VIII.
- Happiness: APL identifies lower nose and upper lip corresponding to patch learning for AU 6 and AU 12 separately. Similar results are shown for fear, sadness, and surprise.

E. Holistic Expression Recognition

As discussed in Sec. IV, JPML can be extended for learning common and specific patches, termed as CS-APL, for holistic facial expressions [21], [43]. We used SPAMS[3] to obtain a solution to patch learning, and 10-fold cross validation for model selection. The original SPAMS toolbox only allows for group-sparsity defined within each column. Recall $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ represents the classification matrix for L expressions, where $\mathbf{w}_\ell$ is the model learned for the ℓ-th expression. To cope with the group-sparsity across columns in Eq. (16), we reformulate the patch learning as follows.

[3] http://spams-devel.gforge.inria.fr/

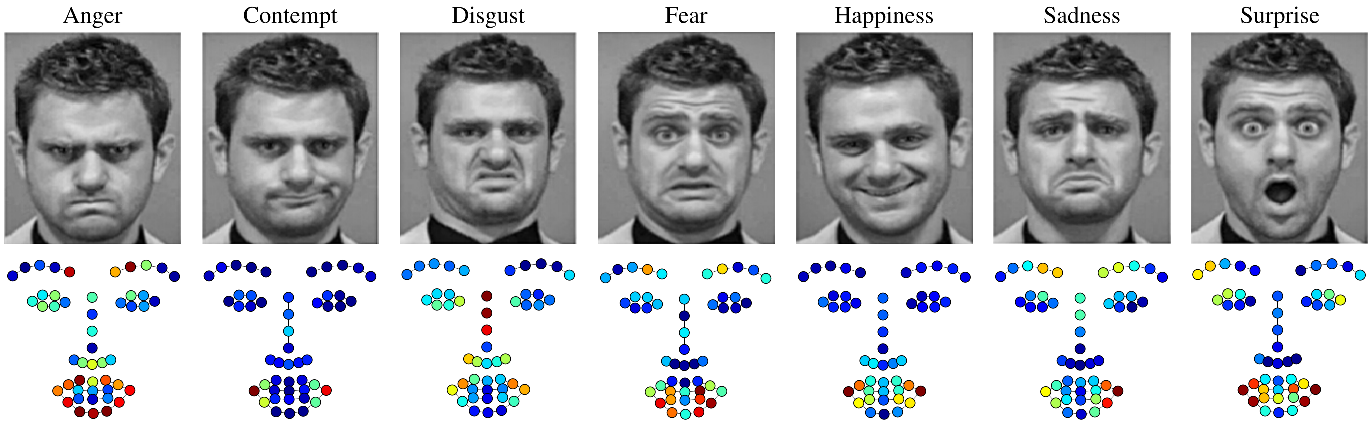


Fig. 10. Visualization of patch importance learned for 7 holistic expressions on the CK+ dataset [22].

TABLE VIII

AU COMPOSITION RULE FOR HOLISTIC FACIAL EXPRESSIONS

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
AUs	4+5+ 7+23	9+15+16	1+2+4+ 5+20+26	6+12+25	1+4+15	1+2+ 5+26

First, we rewrite the original model $\mathbf{W} \in \mathbb{R}^{D\times L}$ into the block diagonal matrix $\overline{\mathbf{W}} \in \mathbb{R}^{DL\times L}$ as:

$$\overline{\mathbf{W}} = \begin{bmatrix} \mathbf{w}_1 & 0 & 0 & 0 \\ 0 & \mathbf{w}_2 & 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{w}_L \end{bmatrix}. \tag{17}$$

Due to linearity of our model, the multi-label output $\mathbf{y} \in \mathbb{R}^L$ for L expressions can be equivalently expressed as:

$$\mathbf{y} = \mathbf{W}^\top \mathbf{x} = \overline{\mathbf{W}}^\top [\underbrace{\mathbf{x}^\top, \ldots, \mathbf{x}^\top}_{L \text{ duplicates}}]^\top. \tag{18}$$

Using (17) and data duplicates, suppose we obtain the solution from SPAMS as:

$$\mathbf{W}^* = \begin{bmatrix} \mathbf{w}^*_{11} & \mathbf{w}^*_{21} & \cdots & \mathbf{w}^*_{L1} \\ \mathbf{w}^*_{12} & \mathbf{w}^*_{22} & & \mathbf{w}^*_{L2} \\ \vdots & & \ddots & \vdots \\ \mathbf{w}^*_{1L} & \mathbf{w}^*_{2L} & \cdots & \mathbf{w}^*_{LL} \end{bmatrix}. \tag{19}$$

Although $\mathbf{W}^*$ is not in the block diagonal form of desire, $\mathbf{W}^*$ and $\overline{\mathbf{W}}$ can be shown interchangeable by observing:

$$\mathbf{W}^{*\top} \left[\underbrace{\mathbf{x}^\top, \ldots, \mathbf{x}^\top}_{L \text{ duplicates}} \right]^\top = \left[\left(\sum_{\ell=1}^{L} \mathbf{w}^*_{1\ell} \right)^\top \mathbf{x}, \ldots, \left(\sum_{\ell=1}^{L} \mathbf{w}^*_{L\ell} \right)^\top \mathbf{x} \right]^\top. \tag{20}$$

Using (18) and letting $\mathbf{w}_i = \sum_{\ell=1}^{L} \mathbf{w}^*_{i\ell}$, $\mathbf{W}^*$ can be written into $\overline{\mathbf{W}}$ by summing to the diagonals and setting off-diagonals to zeros, showing that the solution to $\mathbf{W}$ is equivalently obtained using SPAMS.

TABLE IX

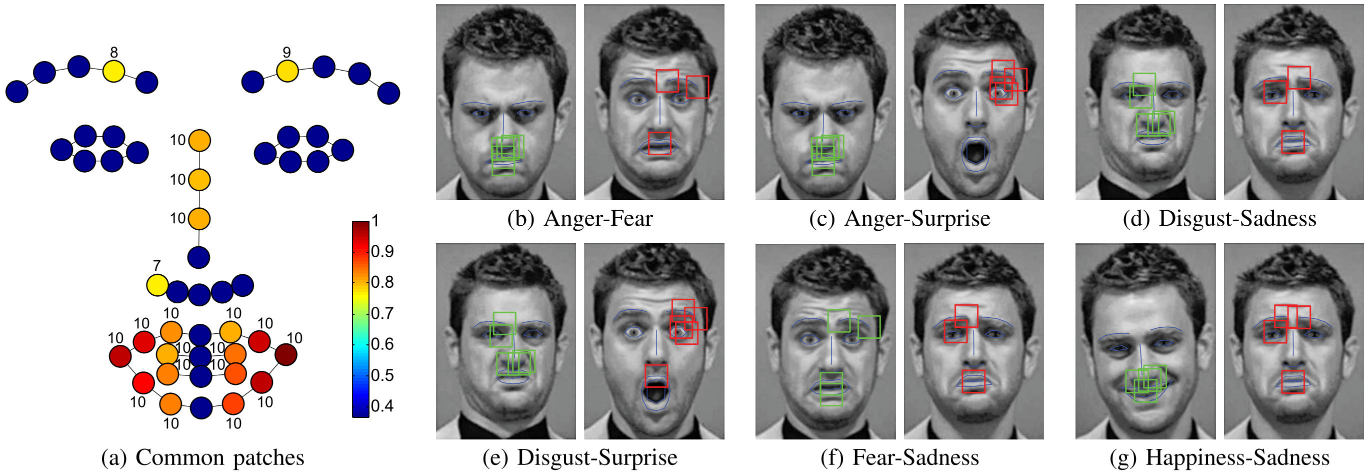
ACCURACY ON THE CK+ DATASET. BRACKETED NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

Expressions	LSVM [12]	FDM [21]	C-PL [43]	CS-PL [43]	APL	CS-APL
Angry	**[97.6]**	–	65.6	71.4	**97.4**	96.1
Disgust	**[99.8]**	–	92.7	95.3	99.5	**99.6**
Fear	**[98.2]**	–	79.0	81.1	**97.3**	**[98.2]**
Happiness	**[99.5]**	–	94.7	95.4	99.3	**99.4**
Sadness	**98.6**	–	87.7	88.0	98.4	**[99.1]**
Surprise	98.7	–	**98.8**	98.3	98.7	**[99.0]**
Avg.	**[98.7]**	97.7	86.5	88.3	98.4	**98.6**

In this experiment, we used CK+ [22] with six holistic expressions. For learning common patches, we put the 128-dimensional $\mathbf{w}_i$ across L expressions ($L = 6$) in one group, thus 49 groups can be obtained. For learning specific patches, with the same group setting, pairwise expressions ($L = 2$) were learned at one time. We term this extension for selecting common and specific patches as Common-Specific Active Patch Learning (CS-APL). In this experiment, top rank 20 patches shared across six expressions were selected as common patches. For each expression, specific patches were selected over all C_6^2 pairs of expressions. Then common and specific patches were imported into an SVM to predict each expression. Table IX shows the accuracy of CS-APL. The tuning parameters in SPAMS were set as $\lambda = 10^{-4}$ and #itr= 500.

We compared CS-APL with the state-of-the-art patch learning methods: Common Patch Learning (C-PL) [43], Common-Specific PL (CS-PL) [43] and Feature Disentangling Machine (FDM) [21]. CS-PL performed a two-stage strategy that first learned common patches as C-PL, and then combined face verification to learn specific patches excluding the common ones. FDM is a unified algorithm to disentangle features into three types of patches, *i.e.*, common patches, specific patches, and useless patches. As in FDM, we used CS-APL (Sec. IV) to disentangle features into the three types of patches, and selected common and specific patches for holistic expression recognition. In addition, we implemented APL (as mentioned in Sec. III-B.1) for comparison.

Table IX shows our results. Since CK+ is a relatively clear-cut and posed dataset, linear SVM with our definition of



(a) Common patches

(b) Anger-Fear

(c) Anger-Surprise

(d) Disgust-Sadness

(e) Disgust-Surprise

(f) Fear-Sadness

(g) Happiness-Sadness

Fig. 11. An illustration of selected patches using CS-APL: (a) shows the *common patches* selected across 6 holistic expressions in CK+ dataset using a 10-fold protocol. The numbers indicate the times that each patch was selected in each fold. The warmer the color, the more patch importance JPML learned. (b)–(g) show the *specific patches* selected for each expression pair following [21]. Blue lines represent the shape of tracked landmarks. Green and red rectangles indicate the selected patches for respective expression pair.

patches reaches a satisfactory result. However, as discussed in Sec. III-B.1, a linear SVM treats each feature equally without consideration of local dependencies as patches. The accuracy for each expression of C-PL and CS-PL are given by the paper [43]. As only the average accuracy for all expressions are given in paper [21], accuracy for each expression is marked as "–" in Table IX. In patch learning that treats features in groups, our proposed common/specific patch learning algorithm performs better when comparing with existing methods.

To visualize the common/specific patches learned by the proposed CS-APL, Fig. 11 shows examples of learned patched by the proposed CS-APL. Fig. 11(a) shows the common patches learned using a 10-fold experiments. There are great overlaps between different fold experiments (almost 10 for all selected patches). It implies that our algorithm is robust to the selection of the training set. The selected common patches are around the areas of mouth and eyebrows, which are consistent with FACS [11], and patches learned by C-PL/CS-PL [43] and FDM [21].

Fig. 11(b)-(g) show a few examples of the learned specific patches. As can be seen, expression-specific patches for the target expression are closely related to FACS [11]. Recall in Table VIII, an expression can be inferred using combinations of AUs. For example, in Fig. 11(d) and (e), the specific patches selected for disgust are located around the nose, which are related to AU 9 (nose wrinkler). For the surprise expression, as shown in in Fig. 11(c) and (e), patches around eyes are the primary AUs, which correspond to AUs 1, 2 and 5 as indicated in Table VIII. Similar results can be found in other expressions. Compared with alternative methods, the proposed CS-APL achieved comparable performance while maintaining interpretable common/specific patches.

F. Discussion

Given the aforementioned experiments, we summarize our observations from the results:

1. JPML outperformed patch learning and multi-label learning in both within- (Sec. V-B) and cross-dataset (Sec. V-C) scenarios, suggesting that jointly modeling dependencies between features and AUs can improve performance.
2. The patch importance learned by JPML (Fig. 8) showed visually similar distributions for AUs with positive correlations (Table I). Results of patch learning methods and JPML suggest that learning with patches is more advantageous than with individual features for AU detection.
3. Holistic expressions can be decomposed into common and specific patches (Fig. 11), and competitive performance can be obtained using only these patches (Table IX), suggesting that patches contribute unequally to expression recognition.

VI. Conclusion

We have presented a Joint Patch and Multi-label Learning (JPML) framework for facial AU and holistic expression recognition. JPML jointly learns a discriminative multi-label classifier and patch importance for different AUs. To model the dependencies between AUs, we statistically derived two types of AU relations, namely *positive correlations* and *negative competitions*, from more than 350,000 annotated video frames. With slight modification, we showed that JPML can be extended to learn common and specific patches for holistic expression recognition. Extensive experiments demonstrated the effectiveness of JPML on three datasets over alternative methods, including standard feature learning and methods based on either patch learning or multi-label learning alone. Future work includes a non-linear extension and an automatic learning of relations between AUs and expressions. Moreover, instead of using manually-defined patches centered at facial landmarks, directly learning the important regions from raw face images could be another direction.

Acknowledgments

Any opinions, findings, and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank the authors in [41] for providing their codes, and Jayakorn Vongkulbhisal for helpful discussion.

REFERENCES

[1] 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 *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2004, pp. 592–597.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[3] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1–9.

[4] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proc. ICDM*, Dec. 2009, pp. 746–751.

[5] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. CVPR*, 2013, pp. 3515–3522.

[6] J. F. Cohn and F. De la Torre, "Automated face analysis for affective computing," in *The Oxford Handbook of Affective Computing*, 2014, p. 131.

[7] J. F. Cohn and A. Zlochower, *A Computerized Analysis of Facial Expression: Feasibility of Automated Discrimination*. New York, NY, USA: American Psychological Society, Jun. 1995.

[8] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, "IntraFace," in *Proc. AFGR*, 2015, pp. 1–8.

[9] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang, "Facial action unit event detection by cascade of tasks," in *Proc. ICCV*, 2013, pp. 2400–2407.

[10] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci.*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[11] P. Ekman, W. Friesen, and J. C. Hager, "Facial action coding system," in *Human Face*, Salt Lake, 2002.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[13] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, 2007.

[14] S. Jaiswal, B. Martinez, and M. F. Valstar, "Learning to combine local models for facial action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–6.

[15] L. A. Jeni, J. F. Cohn, and F. De la Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proc. Conf. Affective Comput. Intell. Interact.*, Sep. 2013, pp. 245–251.

[16] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. AFGR*, Mar. 2011, pp. 314–321.

[17] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.

[18] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. NIPS*, 2010, pp. 1378–1386.

[19] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 127–141, Apr. 2013.

[20] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image Vis. Comput.*, vol. 24, no. 6, pp. 615–625, 2006.

[21] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine—A novel approach of feature selection and disentangling in facial expression analysis," in *Proc. ECCV*, 2014, pp. 151–166.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPRW*, Jun. 2010, pp. 94–101.

[23] S. Lucey, A. B. Ashraf, and J. Cohn, "Investigating spontaneous facial action recognition through AAM representations of the face," *Face Recognit.*, vol. 32, no. 11, pp. 275–286, 2010.

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009, pp. 689–696.

[25] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[26] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[27] M. A. Sayette *et al.*, "Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding," *Psychol. Sci.*, vol. 23, no. 8, pp. 869–878, 2012.

[28] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2140–2152, Jul. 2015.

[29] M. H. Siddiqi, R. Ali, A. M. Khan, Y. T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1386–1398, Apr. 2015.

[30] S. Taheri, Q. Qiu, and R. Chellappa, "Structure-preserving sparse decomposition for facial expression analysis," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3590–3603, Aug. 2014.

[31] U. Tariq *et al.*, "Emotion recognition from an ensemble of features," in *Proc. AFGR*, Mar. 2011, pp. 872–877.

[32] Y. Tong and Q. Ji, "Learning Bayesian networks with qualitative constraints," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[33] Y. Tong, W. Liao, and Q. Ji, "Inferring facial action units with causal relations," in *Proc. CVPR*, 2006, pp. 1623–1630.

[34] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. ICCV*, Dec. 2013, pp. 3304–3311.

[35] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[36] S. Zafeiriou and I. Pitas, "Discriminant graph structures for facial expression recognition," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1528–1540, Dec. 2008.

[37] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong, "Confidence preserving machine for facial action unit detection," in *Proc. ICCV*, 2015, pp. 3622–3630.

[38] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *Proc. CVPR*, 2008, pp. 1–8.

[39] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas, "Automatic image annotation using group sparsity," in *Proc. CVPR*, Jun. 2010, pp. 3312–3319.

[40] X. Zhang *et al.*, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. Autom. Face Gesture Recognit. Workshop*, Apr. 2013, pp. 1–6.

[41] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn, "A lp-norm MTMKL framework for simultaneous detection of multiple facial action units," in *Proc. WACV*, 2014, pp. 1104–1111.

[42] 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 *Proc. CVPR*, 2015, pp. 2207–2216.

[43] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2014.

[44] Y. Zhou, R. Jin, and S. C. H. Hoi, "Exclusive lasso for multi-task feature selection," in *Proc. AISTATS*, 2010, pp. 988–995.

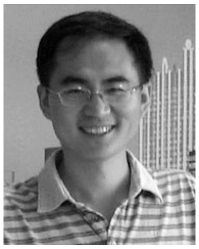
[45] Y. Zhu, F. De la Torre, J. F. Cohn, and Y. J. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 79–91, Apr. 2011.

Kaili Zhao is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. She is involved in research on facial expression analysis.

Wen-Sheng Chu (S'13) is currently pursuing the Ph.D. degree with the Robotics Institute, Carnegie Mellon University. His research interests lie in the development and use of machine learning techniques for computer vision problems. He is a member of the Phi Tau Phi Scholastic Honor Society.

Fernando De la Torre received the B.Sc. degree in telecommunications, and the M.Sc. and Ph.D. degrees in electronics engineering from the La Salle School of Engineering, Ramon Llull University, Barcelona, Spain, in 1994, 1996, and 2002, respectively. He is currently an Associate Research Professor with the Robotics Institute, Carnegie Mellon University. He is currently Directs the Component Analysis Laboratory and the Human Sensing Laboratory, Carnegie Mellon University. He has authored over 150 publications in referred journals and conferences. His research interests are in the fields of computer vision and machine learning. He is an Associate Editor of the IEEE TPAMI. He has organized and co-organized several workshops and has given tutorials at international conferences on component analysis.

Jeffrey F. Cohn is currently a Professor of Psychology and Psychiatry with the University of Pittsburgh and an Adjunct Professor of Computer Science with the Robotics Institute, Carnegie Mellon University. He leads interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of facial expression and prosody, and applies those tools to research in human emotion, social development, non-verbal communication, psychopathology, and biomedicine. He has served as the Co-Chair of the 2008 IEEE International Conference on Automatic Face and Gesture Recognition, the 2009 International Conference on Affective Computing and Intelligent Interaction, the Steering Committee for IEEE International Conference on Automatic Face and Gesture Recognition, and the 2014 International Conference on Multimodal Interfaces. He has Co-Edited special issues of the *Journal of Image and Vision Computing*. He is a Co-Editor of the IEEE TRANSACTIONS IN AFFECTIVE COMPUTING.



Honggang Zhang (SM'12) received the B.S. degree from the Department of Electrical Engineering, Shandong University, in 1996, the master's and Ph.D. degrees from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT), in 1999 and 2003, respectively. He was a Visiting Scholar with the School of Computer Science, Carnegie Mellon University, from 2007 to 2008. He is currently an Associate Professor and the Director of Web Search Center with BUPT. He published more than 30 papers on TPAMI, SCIENCE, Machine Vision and Applications, AAAI, ICPR, ICIP. His research interests include image retrieval, computer vision, and pattern recognition.