Joint Patch and Multi-label Learning for Facial Action Unit Detection

Kaili Zhao[1] Wen-Sheng Chu[2] Fernando De la Torre[2] Jeffrey F. Cohn[2,3] Honggang Zhang[1]
[1]School of Comm. and Info. Engineering, Beijing University of Posts and Telecom., Beijing China. [2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213. [3]Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260.

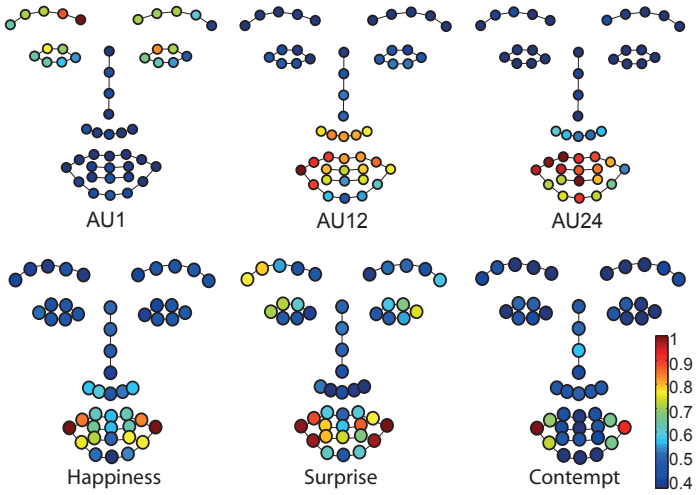

Figure 1: Patch importance for AUs and basic emotions learned by JPML.

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)

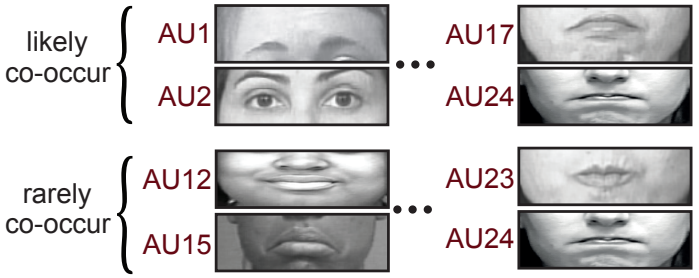

Figure 2: AU relations discovered from >350,000 frames and used in JPML.

To make possible more efficient use of Facial Action Coding System (FACS), computer vision has been devoted to the realization of automatic AU coding. However, two critical problems remain: (1) *patch learning (PL)*: how to exploit local dependencies between features, and (2) *multi-label learning (ML)*: how to model strong correlations between AUs. By modeling features within local patches informed by FACS, it is possible to give greater weights to informative regions and to reduce a large number of correlated features to achieve efficient learning. Similarly, just as features among patches have constrains, AUs have constrains as well. Multi-label learning builds upon the evidence that one AU increases or decreases the likelihood of others.

Joint patch and multi-label learning: We show in this paper how PL and ML can be addressed with one stone, and gradually improve each other. In particular, we propose a framework, termed Joint Patch and Multi-label Learning (JPML), which attempts to model dependencies among both features and AUs. JPML finds a multi-label classification matrix $\mathbf{W}$ constrained with group-wise sparsity and label relations:

$$\min_{\mathbf{W}} L(\mathbf{W},\mathcal{D}) + \alpha\Omega(\mathbf{W}) + \Psi(\mathbf{W},\mathbf{X}), \tag{1}$$

where $L(\mathbf{W},\mathcal{D})$ is the loss function (we used logistic loss), $\Omega(\mathbf{W})$ is the *patch regularizer* that enforces sparse rows of $\mathbf{W}$ as *groups*, $\Psi(\mathbf{W},\mathbf{X})$ is a *relational regularizer* that constrains predictions on $\mathbf{X}$ with AU relations. Problem (1) involves two ends: identify a discriminative subset of patches for each AU (*patch learning*), and incorporate AU relations into model learning (*multi-label learning*).

Patch Learning: To address the regional appearance changes on AUs, we define a group-wise sparsity on the classification matrix $\mathbf{W}$. This paper exploits *landmark patches* that are centered at facial landmarks, and 128-D SIFT descriptors are extracted around each patch. Given the structural nature of our problem, we split the rows of $\mathbf{W}$ into non-overlapping groups, where each corresponds to a *patch* and associates with 128 rows in $\mathbf{W}$. $\Omega(\mathbf{W}) = \sum_{\ell=1}^{L}\sum_{p=1}^{49}\|\mathbf{w}_\ell^p\|_2$ is the *patch regularizer* in problem (1), 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. This grouping encourages a sparse selection of patches by jointly setting 128 values belonging to the same group to zero. As shown in Fig. 1, patch learning offers a better interpretation of important patches corresponding to particular AUs and basic emotions.

Multi-label Learning: Unlike most studies that derive AU relations from domain knowledge, this paper statistically explores AU relations among more than 350,000 frames from CK+ [2], GFT [3], and BP4D [4] datasets. Fig. 2 shows the relations, defined as positive correlation and negative competition, discovered and used in our study. *E.g.*, AUs (6, 12) is a strong positive correlations because they co-occur frequently to describe a Duchenne smile. AUs (12, 15) bears negative competitions because of their negative influences on each other (coincide with the findings in [1]). To incorporate the discovered AU relations, we introduce the *relational regularizer* as:

$$\Psi(\mathbf{W},\mathbf{X}) = \beta_1 PC(\mathbf{W},\mathbf{X},\mathcal{P}) + \beta_2 NC(\mathbf{W},\mathbf{X},\mathcal{N}), \tag{2}$$

where β_1 and β_2 are tradeoff coefficients. $PC(\cdot,\cdot,\cdot)$ and $NC(\cdot,\cdot,\cdot)$ capture the AU relations of positive correlations and negative competitions:

$$PC(\mathbf{W},\mathbf{X},\mathcal{P}) = \frac{1}{2}\sum_{(i,j)\in\mathcal{P}} \gamma_{ij}\|\mathbf{w}_i^\top\mathbf{X} - \mathbf{w}_j^\top\mathbf{X}\|_2^2, \tag{3}$$

$$NC(\mathbf{W},\mathbf{X},\mathcal{N}) = \sum_{i=1}^{N}\sum_{n=1}^{|\mathcal{N}|}\left(\sum_{j\in\mathcal{N}_n}\left|\mathbf{w}_j^\top\mathbf{x}_i\right|\right)^2. \tag{4}$$

Because $\Omega(\mathbf{W})$ and $\Psi(\mathbf{W},\mathbf{X})$ constrain on $\mathbf{W}$ differently, Problem (1) cannot be solved directly. We rewrite Problem (1) by introducing auxiliary variables $\mathbf{W}_1, \mathbf{W}_2$, and then jointly optimize $\mathbf{W}_1$ and $\mathbf{W}_2$ using ADMM.

Results: We compared to a number of baselines and the state-of-the-art patch learning and multi-label learning algorithms. In four of five comparisons on three diverse datasets, CK+, GFT, and BP4D, JPML produced the highest average F1 scores. In no cases, competitive approaches exceed our patch learning and JPML. This suggests that our patch-based approach is more powerful, and further boost the performance with additional ML.

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

[2] 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 *CVPRW*, 2010.

[3] M. A. Sayette, K. G. Creswell, J. D. Dimoff, C. E. Fairbairn, J. F. Cohn, B. W. Heckman, T. R. Kirchner, J. M. Levine, and R. L. Moreland. Alcohol and group formation a multimodal investigation of the effects of alcohol on emotion and social bonding. *Psychological science*, 2012.

[4] X. Zhang, L. Yin, J. F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *AFGRW*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.