Deep Region and Multi-label Learning for Facial Action Unit Detection

Kaili Zhao[1] Wen-Sheng Chu[2] Honggang Zhang[1]
[1]School of Comm. and Info. Engineering, Beijing University of Posts and Telecom., Beijing China
[2]Robotics Institute, Carnegie Mellon University, USA

Abstract

*Region learning (RL) and multi-label learning (ML) have recently attracted increasing attentions in the field of facial Action Unit (AU) detection. Knowing that AUs are active on sparse facial regions, RL aims to identify these regions for a better specificity. On the other hand, a strong statistical evidence of AU correlations suggests that ML is a natural way to model the detection task. In this paper, we propose Deep Region and Multi-label Learning (DRML), a unified deep network that simultaneously addresses these two problems. One crucial aspect in DRML is a novel region layer that uses feed-forward functions to induce important facial regions, forcing the learned weights to capture structural information of the face. Our region layer serves as an alternative design between locally connected layers (*i.e.*, confined kernels to individual pixels) and conventional convolution layers (*i.e.*, shared kernels across an entire image). Unlike previous studies that solve RL and ML alternately, DRML by construction addresses both problems, allowing the two seemingly irrelevant problems to interact more directly. The complete network is end-to-end trainable, and automatically learns representations robust to variations inherent within a local region. Experiments on BP4D and DISFA benchmarks show that DRML performs the highest average F1-score and AUC within and across datasets in comparison with alternative methods.*

1. Introduction

The face reveals thoughts and feelings. Facial expressions, in particular, tell a person's internal states, psychopathology, and social behavior. Facial Action Unit (AU) detection plays a fundamental role in describing comprehensive facial expressions, and has become an important problem in computer vision. In automated facial AU detection, two problems have attracted an increasing attention: *Region Learning (RL)* and *Multi-label Learning (ML)*. Given the definition that an AU is active only on sparse facial regions, RL aims at identifying specific regions to improve detection performance. For example, AU 12 is referred as *lip corner puller*, and by definition is identified only around the region of lip corners. On the other hand, there has been strong statistics showing evidence of correlations between AUs [32, 35]. For instance, AUs 6 and 12 have been observed to frequently co-occur in a Duchenne smile. Building upon these AU correlations, ML attempts to jointly learn multiple AUs as one classification problem. However, it remains unclear how these two problems can interact with each other and jointly be solved.

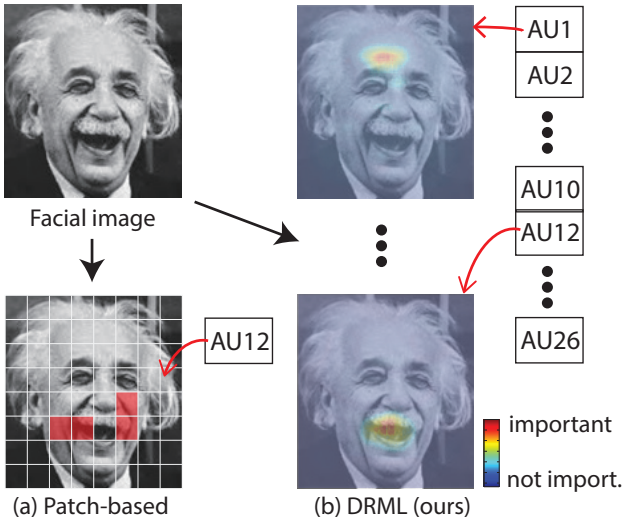

Figure 1. An illustration of (a) a conventional patch-based method, and (b) the proposed DRML. DRML by construction models both important facial regions and relations among multiple AUs, showing a better capability of localization and classification.

Recent work on *patch learning* is a particular example of RL. Conventional patch-based methods divide face images into uniform patches, as shown in Fig. 1(a), and then model the importance for each patch as the magnitude of corresponding model parameters (*e.g.*, [17, 18, 39]). In general, higher importance implies higher relevance for such patches to a particular AU. The selected patches, due to their spatial dependencies, are shown more effective and robust to noise than individual feature values. Nevertheless, the patches are manually defined and the majority of existing work ignores the relationship among AUs. More

recently, multi-label learning showed possibilities of utilizing such AU correlations, *e.g.*, [6, 38]. These works derive AU correlations from FACS heuristics [5] or the statistics from ground truth labels, and then plug the AU correlations into learning, encouraging AUs with high correlation to co-occur more frequently. However, these derived AU correlations can be biased due to coder's subjectiveness or vary from one dataset to another.

In this paper, we propose Deep Region and Multi-label Learning (DRML), a design of neural network that addresses the above issues by construction. Fig. 1 illustrates our main idea. Instead of learning importance on uniform facial grids as shown in Fig. 1(a), DRML propagates contributing value from higher perceptive fields to lower ones. As a result, the more influential "regions" can be discovered, as shown in Fig. 1(b). Due to the multi-label nature of the network, RL and ML can naturally interact with each other within the network, rather than being solved sequentially [39] or alternatively [38]. In addition, we introduce a new region layer that serves as an alternative design between locally connected layers (*i.e.*, confined kernels to individual pixels) and conventional convolution layers (*i.e.*, shared kernels across an entire image). The final network is end-to-end trainable, and converges faster with better learned AU relations than alternative models.

2. Related Work

Automated facial AU detection has been a vital research field for objectively describing facial actions. To tackle AU detection under complex conditions, a majority of studies have been devoted to various features [1, 6, 13, 15, 19] and classifiers [2, 11, 30, 34, 36, 38]. This study is motivated by convolutional neural networks (CNN), and closely related to region learning (RL) and multi-label learning (ML) for AU detection. Below we review each in turn.

Region learning (RL): Conventional methods for AU detection utilize geometric features [4, 19], appearance features [2, 6, 13] or both [3]. Such features are typically quantified as histograms, losing the specificity about facial regions that are critical to indicate existence of AUs [24, 25]. *Region learning* has thus attracted an increasing attention. Zhong *et al*. [39] and Liu *et al*. [18] divided a face image into uniform patches, which are then categorized into common and specific patches to describe different expressions. However, dividing a face image into uniform patches would easily fail on faces with modest or large pose. Taheri *et al*. [27] defined regions for different AUs, and proposed a two-layer group sparsity coding to recover facial expressions using the composition rule of AUs. These regions are pre-defined, and thus can not be learned. Since then, learning the region-AU relation and adaptation to viewpoint changes became a rising demand. Recently, Zhao *et al*. [38] exploited patches centered at facial landmarks, and proposed a multi-label learning framework to infer discriminative patches.

Multi-label learning (ML): Conventional AU detection methods, such AdaBoost [15], GentleBoost [11], or linear SVMs [20], perform detection on individual AUs. On the other hand, given the prior that the occurrence of AUs are strongly correlated [5], multi-label learning (ML) uses the assumption that the correlation exists between labels, so as to improve the detection performance [8]. In addition, AUs are generally unbalanced, *i.e.*, positive samples are outnumbered by negative ones. ML shows potential to address imbalance data learning [35]. For studies considering relationships between AUs, Bayesian Networks (BN) [30] and dynamic BN [32] have been used to learn AU correlations. Recently, Stefanos *et al*. [6] adopted a latent variable CRF to jointly detect multiple AUs. However, they only focus on AUs that co-occur frequently (positive correlation), regardless of the ones that unlikely co-occur (negative competition). Zhao *et al*. [38], instead, statistically derived positive correlations and negative competitions from annotations, and jointly learned multiple AUs using both correlations. Considering pairs or triplets of co-occurring AUs, Zhang *et al*. [36] proposed a multi-task learning approach to learn a common kernel representation that describes the AU correlations.

The aforementioned studies leveraged AU correlations through either FACS heuristic [5] or the statistics from annotations. Such derived AU relations can, thus, be biased due to subjectiveness or data imbalance and could be less transferable. Instead, DRML by construction learns the AU relations and active regions in a unified way. In addition, compared to the closest work Joint Patch and Multi-label Learning (JPML) [38], DRML is an end-to-end trainable and non-linear model, providing a more powerful model to describe AUs under complex conditions. As DRML is inspired by the huge success in convolutional neural networks, we review them below.

Convolutional neural networks (CNNs): CNNs have drastically improved the performance of vision systems, including face verification [10, 28, 29], object detection [14], and video tracking [31]. A standard convolutional layer applies one filter bank to an entire image. For face verification, this spatial stationarity assumption would not hold because different regions have different local statistics for face images [28]. Considering the details of local regions and intra-personal differences, Taigman [28, 29] proposed a locally connected layer (LCN), confining different filters to each pixel location. An LCN, thus, results in a burden of a large number of parameters. For facial expression analysis, Liu *et al*. [17] and Liu *et al*. [16] used convolutional models to learn discriminative local regions for holistic expressions. Liu *et al*. [16] greedily selected AU-aware receptive regions by iteratively learning feature maps with the highest

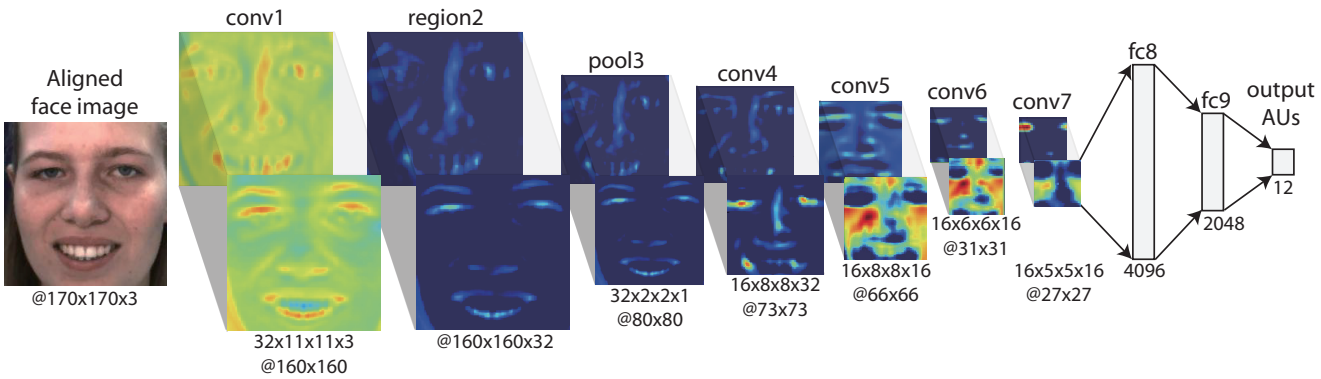

Figure 2. An outline of the proposed DRML architecture. From left to right, a standard convolution layer filtering on an aligned face image, followed by the region layer, one pooling layer and four convolution layers, three fully connected layers, and one multi-label cross-entropy loss layer at the end. Colors illustrate feature maps produced at each layer.

relevance to the target label, then used RBM for classification; CNNs are only used to extract feature maps. Instead of greedily learning local regions, Liu *et al*. [17] proposed an end-to-end framework, utilizing multiple DBNs to learn features with respect to different face regions and strengthening these weak classifiers to top layer in a boosting way. Compared to these models, DRML concentrates on learning discriminative regions. The structural information in local regions is more prominent for AU detection, because AUs depict the local appearance change of faces [5].

Considering dependencies of both local features and AUs correlations, we propose a *region layer* embedded in DRML. The *region layer* confines the same filter for each local region, making the weights in the each region updated individually. Meanwhile, as filters are learned for local regions instead of each pixel [28, 29] or an entire image [14], the updated parameters stand as an alternative between a locally connected layer and a standard convolutional layer. On the other hand, DRML takes both domain knowledge and computation efficiency into account, resulting in an efficient model with comparable performance.

3. Deep Region and Multi-label Learning (DRML)

A common assumption for standard convolutional layers is the shared kernels, or filters, for an entire image. However, for structured objects like faces, such assumption would fail to capture local (and could thus be subtle) appearance changes. To remedy this limitation and make use of AU correlations, we construct a DRML network, with a newly proposed region layer, for multi-label AU detection. In this section, we first discuss the DRML architecture. Then we illustrate the effectiveness of the region layer on learning important regions for different AUs. Finally, we compare similarities and differences between DRML and alternative methods.

3.1. DRML architecture

Fig. 2 shows the outline of the proposed DRML architecture. The principle of designing this network is inspired by the networks for face verification [28, 29]. Because facial appearance changes of AUs are regional and could be subtle, a rule of thumb is to ensure each layer preserves sufficient facial information from its previous layer. Unlike most expression analysis studies that use small face images as input (*e.g*., 48×48 in [22]), we set the input image to 170×170. As shown in Fig. 2, conv7 still maintain a rough face outline to pass to subsequent fully connected layers. Below we detail each layer throughout this network.

The input is an aligned RGB face image, which is then passed to a convolutional layer (conv1) with 32 filters of size 11×11×3. In this paper, we use the notion 32×11×11×3@160×160. The conv1 layer generates 32 feature maps, which are fed into a region layer (region2). Sec. 3.2 provides more details of the region layer, which outputs 32 160×160 feature maps. Following up is a max-pooling layer (pool3), which takes a max operator over 2×2 spatial neighborhoods with a stride 2, separately for each channels of feature maps from the region layer. Because the input face image could obtain modest head pose, the pool3 layer makes the network more robust to small translation errors caused by face alignment. In DRML, we use only one max-pooling layer to avoid losing too much spatial information. The pool3 layer is followed by another four convolutional layers (conv4∼conv7), performing local abstraction as regular CNNs. Finally, two fully connected layers (fc8 and fc9) are placed on top of the conv layers to capture the global correlations across the entire face images. Note that the number of output AUs are relatively small compared to the 1,000 categories in ImageNet [14] or 4,300 identities in DeepFace [28], we keep fc9 as 2048-D instead of 4096-D. The fc9 layer will be extracted as a feature vector for each image. Let the number of AUs be C, the number of samples be N, the ground truth $\mathbf{Y} \in \{-1, 0, 1\}^{N \times C}$,

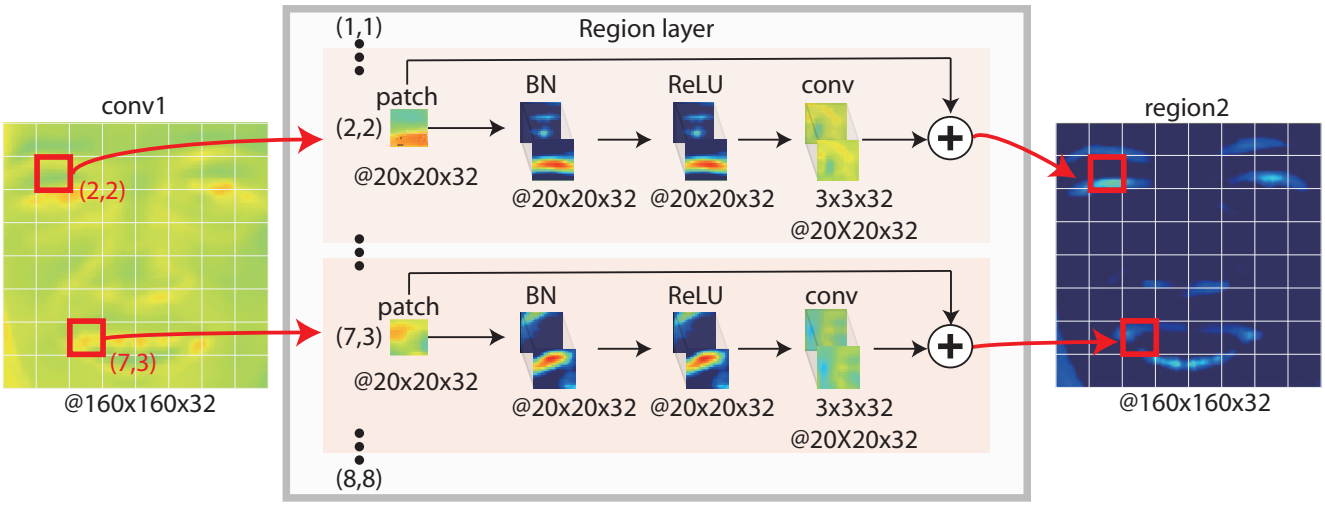

Figure 3. An illustration of the proposed region layer. A feature map is inputed from conv1, and uniformly divided into 8×8 patches. Each 20×20-pixel patch (P_j) is applied with a convolution layer. Re-weight each original patch by adding the convolved one. The output of region layer is a concatenation of all re-weighted patches.

$\mathbf{Y}_{ij}$ indicate the (i, j)-th element of $\mathbf{Y}$, and the predictions $\widehat{\mathbf{Y}} \in \mathbb{R}^{N\times C}$. The output layer was designed as a multi-label sigmoid cross-entropy loss:

$$L(\mathbf{Y}, \widehat{\mathbf{Y}}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\{[\mathbf{Y}_{nc} > 0]\log \widehat{\mathbf{Y}}_{nc} + [\mathbf{Y}_{nc} < 0]\log(1 - \widehat{\mathbf{Y}}_{nc})\},$$

where $[x]$ is an indicator function returning 1 if the statement x is true, and 0 otherwise. It is noteworthy that our resulting model has about 56 million parameters, which is 7% less than AlexNet [14] (60 million) and 53% less than DeepFace [28] (120 million).

3.2. Region layer

One crucial aspect of DRML is to the usage of a *region layer* that captures local appearance changes for different facial regions. Such regional information has shown to provide unique cues to recognize AUs and holistic expressions [24, 25]. Inspired by these works, we designed a *region layer*, whose weights are shared only within a local facial region. Below we interpret its construction and effects on detecting facial AUs.

Most deep learning literature utilize standard convolutional layers to learn image representations (*e.g.*, [14, 16]), and assume weights are shared across an entire image. However, for face images, the spatial stationarity assumption does not hold: Face is more structured than natural images, and, thus, different facial regions follow different local statistics. Motivated by this observation, the authors of DeepFace [28] introduced locally connected layers for face verification. The locally connected layers confine each kernel at each pixel location, resulting in performance that closely approaches humans. However, due to its exhaustive nature, such layers cause a huge number of parameters in the network, *i.e.*, >120 million in DeepFace. For the AU detection task in hand, AU annotations are typically insufficient even for contemporary datasets. For example, there are only ∼140,000 frames in BP4D dataset [37]. Having such a large network could, thus, easily lead to overfitting.

Fig. 3 depicts the proposed region layer, which contains three components: patch clipping, local convolution, and identity addition. The patch clipping component uniformly slices a 160×160 response map into a 8×8 grid. We enumerated different clipping parameters starting from 5×5, and found 8×8 performed the best for our datasets. Each mini-batch is normalized using Batch Normalization (BN), and passed through ReLU [23]. A local convolution component learns to capture local appearance changes, forcing the learned weights in each patch to be updated independently. An identity addition component is introduced along with a "skip connection" from the input patch, which helps avoid vanishing gradient issues during training the network [9]. Imposing the skip connection also simplifies the learning hypothesis: If an input patch contains no useful information for detecting a particular AU, it would be easier to directly forward the patch than learning a filter bank to reduce the patch's effect. As we will see in our experiments, adding this layer helps preserve sparse facial regions activated by particular AUs [5].

What does *region layer* capture for AU detection? Here we illustrate that *region layer* can induce important facial regions for identifying different AUs. Specifically, we adopt a "saliency map" [26] to visualize the regions selected by different models with and without a region layer. The saliency map is image-specific, and computed as the

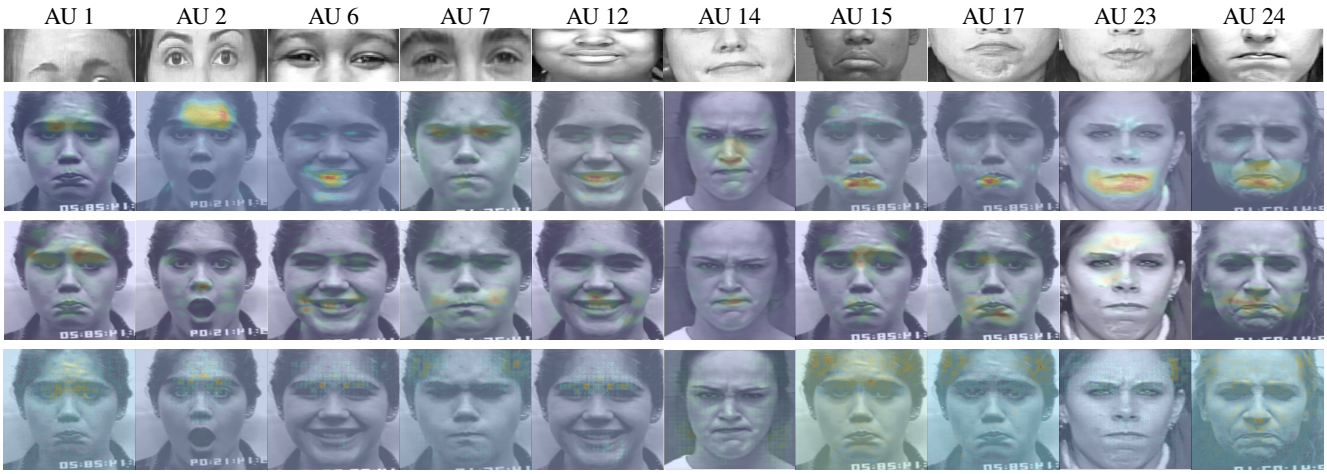

Figure 4. Visualization of AU-specific saliency maps for three networks: DRML (second row), ConvNet (third row), and AlexNet [14] (bottom row). The top row illustrates the appearance of 10 AUs. Colors on each map indicate saliency intensity from low high.

magnitude of per-pixel gradients with respect to a particular AU. We treat such gradient magnitude as the "active region" of a face image. In this way, we are able to investigate the important and active regions for each AU.

To show the specificity of facial regions learned using the region layer, we compare DRML with a standard ConvNet (DRML architecture without the region layer) and the AlexNet [14]. All networks were trained on the BP4D dataset [37] and used the multi-label sigmoid cross-entropy loss. Fig. 4 shows the learned active patches of 10 common AUs. For illustration purpose, the face images were selected manually with apparent AUs from the CK+ dataset [20]. As can be seen, DRML learns a more specific and concentrated regions for the corresponding AUs. Below we summarize our observations:

- **AUs 1, 2**: For AU1, DRML identifies important regions around inner eyebrow, emphasizing the appearance changes by pulling inner eyebrows. On the contrary, ConvNet and AlexNet emphasize some eye regions, but not as concentrated as DRML. For AU2, DRML has a high saliency on the forehead and strong outer brows. The presence with slight saliency on inner brows indicates its likely co-occurrence with AU1. ConvNet marks some lower face and AlexNet fail to concentrate.
- **AUs 6, 7**: For AU6 (check raiser), DRML identifies the center of the mouth due to the strong positive correlation of AU6 and AU12 [38]. For AU7, DRML gives much more importance on eyelids than the other identified regions. ConvNet also identifies mouth regions for AU6, but check regions for AU7. AlexNet fails to identify saliency for AU6 and AU7.
- **AU12**: AU12 depicts lip corner puller, commonly seen in smiley face. DRML concentrates more on the teeth and slight on the eyes and cheek. ConvNet not only identified the mouth but also some chin regions. AlexNet fails to identify meaningful regions for AU12.
- **AU14**: AU14 is dimpler, causing appearance changes of mouth corners. DRML emphasizes the regions around nose; ConvNet regions around mouth. AlexNet fails to identify the subtle appearance changes of AU14.
- **AUs 15, 17**: AU15 and AU17 depict lip corner depressor and chin raiser, which both could cause appearance changes around lower mouth. We observe that DRML is able to concentrate salient regions on lower mouth for AU15 and AU17. ConvNet emphasizes regions around mouth and some regions on the upper face. AlexNet shows saliency over the whole face for AU15 and AU17.
- **AUs 23, 24**: AUs 23 and 24 depict lip tightener and lip pressor. DRML identifies strong saliency around mouth, while ConvNet emphasizes on regions of both mouth and the upper face. AlexNet fails these two AUs.

In all, DRML identifies concentrated and sparse regions than alternative methods. These identified regions also coincide with the important patches in JPML [38]. We found AlexNet consistently fails to identify specific active regions. One reason is the per-pixel contribution of gradient could look like salt-and-pepper noise. Adding further regularization (*e.g.*, [33]) might help the visualization. Recall that ConvNet is a special case of DRML without the region layer. From this perspective, adding the region layer can be regarded as an regularizer that helps reveal the sparse and discriminative regions. We thus infer that the architecture constructed in this paper is more suitable for AU detection.

3.3. Comparison with related work

DRML shares similarities with patch-based methods for AU detection, *i.e.*, Active Patch Learning (APL) [39], JPML [38], AUDN [16], and Boosted DBN (BDBN) [17]. All aims to select a discriminative subset of facial regions for better AU detection. However, they differ in several aspects. Table 1 summarizes these differences.

Table 1. Comparisons between DRML and alternative methods

Methods	ET	ML	LR	NL	OU
APL [39]	×	×	×	×	×
AUDN [16]	×	×	✓	✓	×
BDBN [17]	✓	×	✓	✓	✓
JPML [38]	×	✓	×	×	×
DRML	✓	✓	✓	✓	✓

*__ET__: end-to-end trainable, __ML__: multi-label learning, __LR__: learning representation, __NL__: non-linearity, __OU__: online update.

APL [39] selects patches for different expression by inducing sparsity on "groups", which are defined over uniform patches. Different from expression recognition, AU is a multi-label learning problem. In view of dependencies among features and AUs, JPML [38] jointly learns discriminative patches for multiple AUs. The proposed DRML is inspired by JPML. However, they bear several differences. First, JPML defines AU relations through dataset statistics; DRML by construction learns correlations among AUs. Second, JPML uses manually-crafted feature (*i.e.*, SIFT); DRML learns the features that adapts to multi-label AU detection. Third, JPML learns the PL and ML alternatively; DRML naturally fuses two tasks into one framework, allowing ML and PL to interact more directly. Finally, JPML is linear; DRML stacks non-linear functions that potentially better models the non-linearity of facial AUs.

Learning representations from raw face images is another crucial property of DRML. AUDN [16] and BDBN [17] also have this property for expression recognition. AUDN [16] sequentially combined three modules that respectively learn expression-specific representation, search subset of the representation that best simulates an AU, and concatenate the subset for recognition. However, these three modules are trained independently; DRML is end-to-end trainable. BDBN [17] integrated feature learning, patch selection and classifier construction into one end-to-end trainable framework. Each patch is associated with one DBN. The selection process was done by forming the DBNs as a strong boosted classifier. However, building a network for each patch can be very expensive. Instead, DRML performs the selection through a region layer, containing much smaller units and an identity connection that allows more direct gradient flows.

4. Experiments

4.1. Settings

Datasets: We evaluated DRML on two spontaneous datasets: BP4D [37] and DISFA [21]. For BP4D, we adopted a 3-fold partition to ensure subjects were mutually exclusive in train/val/test sets. For DISFA, we reported results using the best model obtained from BP4D.

(1) BP4D [37] contains 2D and 3D videos of 41 young adults during various emotion inductions while interacting with an experimenter. We used 328 videos (41 participants×8 videos each) with 10 AUs coded, resulting in ∼140,000 valid face images. For each AU, we sampled 100 positive frames and 200 negative frames for each video.

(2) DISFA [21] contains 27 subjects watching video clips, and provides 8 AU annotations with intensities. There were ∼130,000 valid face images. We used the frames with AU intensities with C-level or higher as positive samples, and the rest as negative ones. To be consistent with the 8-video setting of BP4D, we sampled 800 positive frames and 1600 negative frames for each video.

Metrics: The performance was evaluated on two common frame-based metrics: F1-frame and AUC. F1-frame is the harmonic mean of precision and recall, and widely used in AU detection. AUC quantifies the relation between true and false positives. For each method, we computed average metrics over all AUs (denoted as *Avg.*).

Implementation: We registered face images to 200×200 using similarity transform [34, 38]. Each face was randomly cropped into 170×170, or horizontally mirrored for data augmentation. All models were initialized with learning rate of 0.001, which was further reduced after 8000 iterations. A momentum of 0.9 and weight decay of 0.0005 was used. All implementations were based on the Caffe toolbox [12] with modifications to support the region layer and multi-label cross-entropy loss. All experiments were performed on one NVIDIA Tesla K40c GPU. Our implementation is available online[1].

Comparative methods: We compared DRML to alternative methods, including a baseline linear SVM (LSVM) [7], a patch-learning method APL [39], and the state-of-the-art Joint Patch and Multi-label Learning (JPML) [38]. For baseline networks, we compared to AlexNet [14], ConvNet (DRML excluding the region layer), and LCN (ConvNet with locally connected layer [28]). These alternative approaches were picked to answer several questions: (1) whether the learned features are more descriptive than hand-crafted ones, (2) whether using AU correlations and/or region layer improves the performance of AU detection, (3) whether, compared to JPML, the proposed DRML provides a more direct and effective way to jointly learn RL and ML. We excluded JPML in the DISFA experiments due to the lack of reported AU correlations; instead we only reported APL. Below we discuss the results.

4.2. Results

Tables 2 and 3 show the results of 12 AUs for BP4D and 8 AUs for DISFA, respectively. Below we discuss the results from five perspectives: feature representation, multi-label learning, effects of region layer, joint learning of regions and multi-label, and running time.

[1] https://github.com/zkl20061823

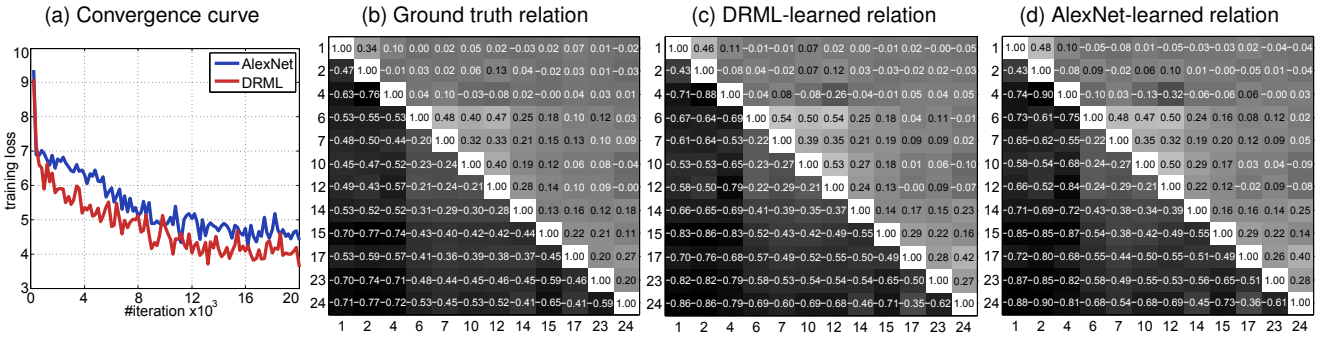

Figure 5. Comparison between DRML and AlexNet [14]: (a) training loss, (b)-(d) relation tables, where each entry (i, j) is computed as the coefficient correlation between the i-th and the j-th AUs.

Feature representation: This paragraph discusses the benefits of the learned features. Comparing the results of AlexNet and LSVM in Table 2, 9 out of 12 AUs in F1-frame and 12 out of 12 AUs in AUC are higher for AlexNet. The improvement of AlexNet became larger on DISFA, showing its better generalizability in a cross-dataset scenario. Recall that the results on DISFA were reported using a cross-dataset protocol, *i.e.*, we selected the best performing model from BP4D to report the results. As shown in Table 3, compared to LSVM, AlexNet achieved about 2% and 13% higher F1-frame and AUC. It is worth noticing that the feature dimensions for LSVM, AlexNet, LCN, DRML are 6272 (128 SIFT features for 49 landmarks), 4096, 2048, and 2048, respectively. In fact, even though the learned features are of lower dimension, more than 40% of learned features for AlexNet, LCN, and DRML, are zeros. We can infer that the learned features, compared to the best performing handcrafted SIFT feature [40], capture more discriminative and sparse characteristics for detecting AUs.

Multi-label learning: Multi-label learning could improve AU detection by taking AU correlations into account. In our experiments, this improvement is more obvious for highly skewed AUs, given the skewness factor defined as the ratio of the number of negative samples to the number of positive samples. Take the BP4D dataset for example. The skewness for frequently occurring AUs (10,12) are (0.7, 0.8) respectively; for infrequently occurring AUs (1,2,23,24), their skewness are (3.8, 4.9, 5.0, 5.5). Compared to the baseline LSVM results, the improvement of methods using multi-label learning (*i.e.*, AlexNet, ConvNet, LCN, and DRML) improves more on AUs with larger skewness. For instance, for both the F1-frame and AUC, the performance of these methods on AUs (1,2,23,24) are 1.5 to 1.8 times higher than the baseline LSVM. That being said, when the training data are relatively rare, multi-label learning helps reduce the effects of the imbalance nature for AU detection. Regardless of the overall improvement across 12 AUs, it is noticeable that, for AUs (10,12), the LSVM baseline achieved satisfactory performance. One possible explanation is that AUs 10 and 12 have relatively abundant training samples compared to other AUs.

Region layer: This paragraph discusses the effectiveness of the region layer. Observing the results of ConvNet and LCN in Table 2, LCN reached higher F1-frame in 11 out of 12 AUs and higher AUC in 7 out of 12 AUs. It validates the observation that LCN learns more discriminative information of face regions than a standard convolutional layer as ConvNet [28]. In BP4D, DRML outperformed LCN in 6 out 12 AUs for F1-Frame, and 8 out 12 AUs for AUC. In DISFA (Table 3), on average, DRML outperformed LCN with 11.3% higher in F1-frame and 11.7% higher in AUC. This justifies that, compared to LCN applied to individual pixels, the region layer better expresses the structural information in local facial regions. Recall that ConvNet is a special case of DRML without the region layer, we confirmed the effectiveness of DRML. Qualitatively, as shown in Fig. 4, the saliency maps of DRML show better specificity than alternative models. All results suggest that the proposed region layer helps AU detection by considering structural information in facial regions.

Joint learning of regions and multi-label: To better understand the effects of a joint learning framework, we compared the proposed DRML with an AlexNet [14]. Both networks were trained on the BP4D dataset using 12 AUs. Fig. 5 shows the convergence curves and learned relation tables of both models. As shown in Fig. 5(a), DRML convergences faster than AlexNet and obtains lower training loss. Fig. 5(b)-(d) show the table of correlation coefficients between pairwise AUs for ground truth, DRML and AlexNet, respectively. The element-wise Euclidean distance between DRML and ground truth is 0.0068, while it is 0.0077 for AlexNet. This shows that DRML was able to learn AU relations close to ground truth statistics. In addition, we compared DRML with the state-of-the-art JPML [38]. Note that one difference is that JPML [38] reported their results using a 10-split partition, while, for fairness, we implemented JPML using a 3-split partition. Observing the results in Table 2, on average, DRML achieved 5.0% higher in F1-frame and 11.7% higher in AUC. The results suggest that the direct interaction between RL and ML, along with the non-

Table 2. Results on the BP4D dataset. Bracketed and bold numbers indicate the best performance; bold numbers indicate the second best.

	F1-frame						AUC					
AU	LSVM	JPML	AlexNet	ConvNet	LCN	DRML	LSVM	JPML	AlexNet	ConvNet	LCN	DRML
1	23.2	32.6	27.0	**40.4**	**[45.0]**	36.4	20.7	40.7	34.9	49.4	**51.9**	**[55.7]**
2	22.8	25.6	25.5	**[46.1]**	41.2	**41.8**	17.7	42.1	25.8	**[51.3]**	50.9	**[54.5]**
4	23.1	37.4	31.9	**42.8**	42.3	**[43.0]**	22.9	46.2	36.1	47.4	**53.6**	**[58.8]**
6	27.2	42.3	51.4	51.8	**[58.6]**	**55.0**	20.3	40.0	48.3	52.2	**53.2**	**[56.6]**
7	47.1	50.5	**55.4**	54.3	52.8	**[67.0]**	44.8	50.0	54.3	**[64.8]**	**63.7**	61.0
10	**[77.2]**	**72.2**	52.8	54.0	54.0	66.3	**73.4**	**[75.2]**	54.3	61.4	62.4	53.6
12	63.7	**[74.1]**	49.0	61.0	54.7	**65.8**	55.3	**60.5**	50.0	60.2	**[61.6]**	60.8
14	**64.3**	**[65.7]**	51.7	56.7	59.9	54.1	46.8	**53.6**	47.7	29.8	**[58.8]**	57.0
15	18.4	38.1	25.5	**[44.1]**	**36.1**	33.2	18.3	50.1	34.9	**50.6**	49.9	**[56.2]**
17	33.0	40.0	41.4	38.3	**46.6**	**[48.0]**	36.4	42.5	48.5	**[53.5]**	48.4	**50.0**
23	19.4	30.4	26.1	**[41.8]**	**33.2**	31.7	19.2	**51.9**	40.5	49.5	50.3	**[53.9]**
24	20.7	**[42.3]**	23.5	32.8	**35.3**	30.0	11.7	**53.2**	31.7	52.5	47.7	**[53.9]**
Avg.	35.3	45.9	38.4	**47.0**	46.6	**[48.3]**	32.2	50.5	42.2	51.8	**54.4**	**[56.0]**

Table 3. Results on the DISFA dataset. Bracketed and bold numbers indicate the best performance; bold numbers indicate the second best.

	F1-frame						AUC					
AU	LSVM	APL	AlexNet	ConvNet	LCN	DRML	LSVM	APL	AlexNet	ConvNet	LCN	DRML
1	10.8	11.4	12.0	11.7	**12.8**	**[17.3]**	21.6	32.7	**47.8**	44.2	44.1	**[53.3]**
2	10.0	12.0	11.6	12.0	**12.0**	**[17.7]**	15.8	27.8	52.1	37.3	**52.4**	**[53.2]**
4	21.8	**30.1**	27.6	28.9	29.7	**[37.4]**	17.2	37.9	44.0	**47.9**	47.7	**[60.0]**
6	15.7	12.4	22.6	21.4	**23.1**	**[29.0]**	8.7	13.6	**44.3**	38.5	39.7	**[54.9]**
9	11.5	10.1	11.5	11.5	**[12.4]**	10.7	15.0	**[64.4]**	48.7	49.5	40.2	**51.5**
12	**[70.4]**	**65.9**	31.1	31.0	26.4	37.7	**93.8**	**[94.2]**	55.3	54.8	54.7	54.6
25	12.0	21.4	**44.4**	40.7	**[46.2]**	38.5	3.4	**[50.4]**	**50.2**	48.4	48.6	45.6
26	22.1	26.9	**28.2**	27.7	**[30.0]**	20.1	20.1	**[47.1]**	45.8	45.8	**47.0**	45.3
Avg.	21.8	23.8	23.6	23.1	**24.0**	**[26.7]**	27.5	46.0	**49.1**	45.8	46.8	**[52.3]**

Table 4. Running time (ms) of all alternative networks

Time	**ConvNet**	**AlexNet**	**DRML**	**LCN**
Train	9.7±0.003	5.8±0.005	31.2±0.001	74.7±0.006
Test	3.3±0.002	2.2±0.002	12.1±0.003	34.7±0.008

linearity, bring more advantages to DRML over JPML.

Running time: We evaluated the running speed of DRML and alternative networks using a NVIDIA Tesla K40c GPU. Table 4 shows the execution time (ms) for both training and test phases. Specifically, using the same settings as described in Sec. 4.1, we ran each network for 20 trials over 1,000 iterations, evaluated the running time for each iteration, and then computed the mean and standard deviation over the 20 trials. Because DRML serves an alternative architecture between ConvNet and LCN, both training and test time of DRML falls between them. Note that DRML is significantly faster than LCN, which was proposed for face verification [28]. It is worth noticing that, even ConvNet has slightly smaller number of parameters than AlexNet, the computation complexity could vary, causing the running time of ConvNet slightly larger. In particular, the 11×11 filters in conv1 lead to the major FLOP (multiply-adds) operations.

5. Conclusion

This paper presents Deep Region and Multi-label Learning (DRML) for facial AU detection. DRML is a unified architecture for AU detection, and allows two seemingly irrelevant tasks, region learning (RL) and multi-label learning (ML), to interact directly. DRML is end-to-end trainable, and able to identify more specific regions for different AUs than conventional patch-based methods. To this end, we introduce a region layer that uses feed-forward functions to capture structural information in different facial regions. Experiments conducted on within- and across-dataset scenarios manifest the effectiveness of DRML. Future work includes imposing group sparsity loss into the objective of DRML to learn sparser facial regions. The proposed region layer introduces potential applications to more structured objects, such as cats, cars, and pedestrians.

Acknowledgement: Kaili Zhao is supported by BUPT Excellent Ph.D. Foundation. We thank Prof. Shiguang Shan for initializing this work, who is with VIPL lab of Chinese Academy of Sciences. Most of this work was done there. We thank Jayakorn Vongkulbhisal for helpful comments.

References

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, 2013.

[3] 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 *ICCV*, 2013.

[4] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111:1454–1462, 2014.

[5] P. Ekman, W. Friesen, and J. Hager. Facs manual. *A Human Face*, 2002.

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

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[8] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[10] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, 2012.

[11] S. Jaiswal, B. Martinez, and M. F. Valstar. Learning to combine local models for facial AU detection. In *AFGR*, 2015.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACMMM*, 2014.

[13] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *AFGR*, 2011.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep CNNs. In *NIPS*, 2012.

[15] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *IVC*, 24(6):615–625, 2006.

[16] M. Liu, S. Li, S. Shan, and X. Chen. AU-aware deep networks for facial expression recognition. In *AFGR*, 2013.

[17] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted DBN. In *CVPR*, 2014.

[18] 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 *ECCV*, 2014.

[19] S. Lucey, A. B. Ashraf, and J. F. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. INTECH Open Access Publisher, 2007.

[20] Lucey *et al.* The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *TAFFC*, 4:151–160, 2013.

[22] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*. 2012.

[23] C. S. S. Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[24] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh. Robust representation and recognition of facial emotions using extreme sparse learning. *TIP*, 24(7):2140–2152, 2015.

[25] 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. *TIP*, 24(4):1386–1398, 2015.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

[27] S. Taheri, Q. Qiu, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *TIP*, 23:3590–3603, 2014.

[28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *CVPR*, 2015.

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

[31] N. Wang, Yeung, and Dit-Yan. Learning a deep compact image representation for visual tracking. In *NIPS*, 2013.

[32] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial AU recognition. In *ICCV*, 2013.

[33] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv:1506.06579*, 2015.

[34] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. In *ICCV*, 2015.

[35] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu. Towards class-imbalance aware multi-label learning. In *IJCAI*, 2015.

[36] X. Zhang and M. H. Mahoor. Task-dependent multi-task multiple kernel learning for facial au detection. *PR*, 2015.

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

[38] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *CVPR*, 2015.

[39] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics*, (99), 2014.

[40] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang. Dynamic cascades with bidirectional bootstrapping for AU detection in spontaneous facial behavior. *TAFFC*, 2(2):79–91, 2011.