Facial Action Units Detection Under Pose Variations Using Deep Regions Learning

Asem M. Ali[†], Islam Alkabbany[†], Amal Farag[‡], Ian Bennett[‡] and Aly Farag[†]
[†] *Computer Vision and Image Processing Laboratory (CVIP Lab.), University of Louisville, Louisville, KY 40292*
[‡] *TSN, Inc., Palo Alto, CA 94303*

***Abstract*—A set of facial Action Units (AUs) is activated due to the movement of facial muscles in response to a person's internal emotion. The activated facial action units appear around sparse regions on the face e.g., the mouth and eyes. This group occurrence of AUs reveals the semantic relationships among them. These relationships should be considered in the detection process of AUs.**

Therefore, we propose an approach that learns these semantic relationships using a multi-label deep learning architecture. The sparse patches of the AUs are used in the proposed approach instead of the whole facial region. To handle these sparse patches, we propose a region-based network instead of the well-known convolutional network and the locally connected network. Moreover, unlike the current approaches, which define facial regions using a uniform grid, the proposed region-based architecture, define patches around facial landmarks. This overcomes the region displacement problem of the uniform grid. Usually, the AUs classification suffers from the high skewness factor problem. To overcome this problem, we use a weighted loss function.

We conduct our experiments on two standard benchmarks: BP4D and FERA17. f_1 score, the area under the ROC (AROC) and the area under the precision-recall curve (APR) are used as performance measures. Compared to the standard convolutional layer, the proposed patch-based layer is more effective in capturing the required structural features of the face and learns the correlations among AUs under pose variations. Also, it has been shown that the proposed approach outperforms the state-of-the-art methods.

1. Introduction

The face is an important tool for nonverbal social communication. Thus analysis of facial movement is an active research topic for behavioral scientists since the work of Darwin in 1872 [1]. Facial Action Coding System (FACS) was presented by Ekman and Friesen [2]. After that FACS became the most used method for measuring these facial movements i.e., Action Units (AUs). Action units have a broad impact on several facial expression-based applications such as human-computer interaction [3] and measuring student's engagement [4].

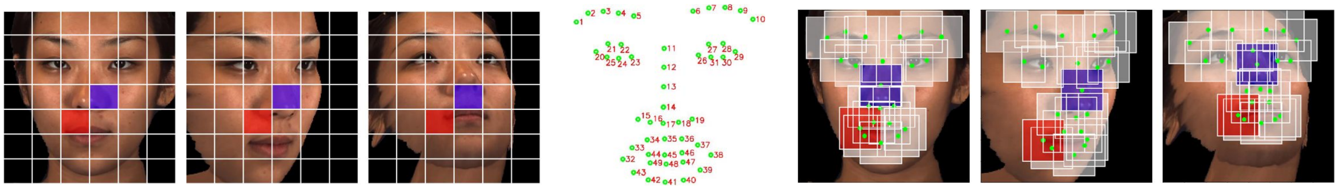
According to a study by G. Duchenne [5], who electrically stimulated facial muscles, movement of the muscles around the mouth, nose and eyes constitute the facial expressions. This reveals the sparse nature of the dominant AUs regions. Therefore, the performance of AUs detectors can be enhanced using region-based signatures. These signatures can be extracted from uniform patches (e.g., [6], [7], [8], [9]) or from patches centered around facial landmarks (e.g., [10], [11]). Instead of directly defining these patches, Li et al. [12] introduced a deep learning-based approach to find important areas and crop these regions of interest. From a psychological point of view, recently, Liu et al. [9] investigated the effect of each facial region on various facial expressions. Similarly, Zhong et al. [7] identified the active facial patches of each facial expression. In the Joint Patch and Multi-label Learning (JPML) approach [10], 49 patches are chosen around facial landmarks. Then these sparse facial patches are used to learn a multi-label classifier. For each action unit, the authors identified the most effective set of those patches.

A single person's emotion activates a set of AUs [13]. As an example, the smile expression simultaneously activates “Lip Corner Puller” and “Cheek Raiser” action units. Therefore, detecting AUs individually (i.e., one-vs-all classification such as SVM [14] and ADABoost [15]) does not exploit these semantic relationships. On the other hand, many researchers (e.g., [6], [10], [11], [16], [17]) investigated the correlations among different action units. To learn these relationships, Tong and Ji [18] used a Bayesian network model and Wang et al. [17] used a restricted Boltzmann machine. In the JPML approach [10], Zhao et al. proposed a multi-label classifier to identify AUs that co-occur frequently and others that unlikely co-occur.

Features that are used in AUs detection can be categorized into: appearance-based features (e.g., SIFT, histogram of gradient (HOG) and Local Binary Pattern (LBP)) [11], [19], geometric-based features [20] or both [21]. The appearance-based features (e.g., a 6272-D SIFT feature vector representing each patch in JPML [10]) are histogram descriptors without any shape information. On the other hand, the geometric-based features ignore any visual information. Recently, features that are learned by deep learning approaches replace these hand crafted features. As an example, a Convolutional Neural Network (CNN) model [22] was proposed to jointly learn dynamic appearance and shape features for facial AUs detection. A Deep Region and Multi-Label learning (DRML) network [6] was proposed to capture local appearance changes for facial regions. A recent CNN-based facial action unit detection approach is EAC-Net [12], which enhances a pre-trained CNN model to learn both feature enhancing and region cropping functions.

Pose variance is one of the main causes that degrade the performance of AUs detector even using the state-of-the-art deep learning approaches. Therefore, the majority of the presented CNN-based models addressed AUs detection for frontal or near-frontal faces. To detect AUs in non-frontal faces, Tosér et al. [23] proposed a deep learning model that tracks the facial fiducial landmark of the individuals and used them to obtain a normalized face. Another cause for the performance degradation is that negative samples predominate the positive ones. This is a common problem for imbalanced large scale multi-label learning frameworks [24], [25].

978-1-5386-0563-9/17/$31.00 ©2017 IEEE



(a) (b) (c)

Figure 1. In the presence of pose, the uniform grid (a) suffers from lack of correspondences (red and blue rectangles) due to displacement and occlusion. To minimizes this lack of correspondence, facial landmarks (b) are used to define a sparse set of patches (c).

To overcome this limitation, Zhang et al. [26] proposed a class-imbalance aware algorithm.

In the proposed work, we exploit both the sparse nature of the dominant AUs regions and semantic relationships among AUs for action units detection. First, to handle pose variations, we define patches around facial landmarks instead of using a uniform grid, which suffers from displacement and occlusion problems as shown in Fig. 1. Then, we propose a new deep region-based neural networks architecture in a multi-label setting to learn both the required features as well as the semantic relationships of AUs. Moreover, we use a weighted loss function to overcome the imbalance problem in multi-label learning.

2. The Proposed Framework

Face alignment is the first step in any facial system. We begin by detecting 49 facial landmarks (see Fig. 1(b)) using our detector [27]. Then, the facial image is aligned by transforming these landmarks to a common space to eliminate the in-plane rotation. Finally, a region of interest is cropped to 200×200 such that the left corner of the right eye becomes the origin of the common space.

Recently, Convolutional Neural Network (CNN) has been presented as an end-to-end framework that performs both feature extraction and classifier training. However, the convolutional layers treat image pixels equally. This spatial stationarity does not hold in faces i.e., structured objects. On the other hand, locally connected layers treat each image pixel differently. But this needs a huge number of parameters to be tuned. To treat each region differently, a recent region-based layer was proposed by Zhao et al., [6]. However, regions are defined using a uniform grid, which is prone to lack of correspondence in the presence of pose as shown in Fig. 1.

Our proposed network architecture overcomes these drawbacks by treating each region differently. Moreover, patches are defined around facial landmarks instead of a uniform grid. We define 22 patches to be 48×48 pixels centered around 22 landmarks out of the 49 landmarks. The 22 overlapped patches were chosen to cover the area of interest in the face as shown in Fig. 1(c). The proposed network architecture, which is inspired by architectures presented in [6] and [28], is shown in Fig. 2. The input to the proposed network is the aligned RGB facial image and its 22 landmarks. First, the image is filtered using 32 filters of size $11 \times 11 \times 3$. This convolutional layer "Conv1" is used to extract a set of low-level features.

Subsequently, 22 patches are extracted from the 32 feature maps (i.e., outputs of "Conv1") around the specified landmarks (which are justified to fit the new size i.e., 190×190) with size of $48 \times 48 \times 32$. Then local features are extracted from each patch by applying five consecutive sets of filters (i.e., "Conv2" - "Conv6") as shown in Fig. 2. In each layer, the number and the size of the filters are the same for each patch but with different weights. As an example, in the convolutional layer "Conv2", there are 22 sets of filters. Each set has 32 filters of the same size $7 \times 7 \times 32$ but different wights. To guarantee the non-linearity in this cascade, an activation function is applied after each layer. Rectified Linear Unit (ReLU) [29] is selected to be the activation function due to its sparse features output. This sparsity is an encouraged behavior for the deep network layer because it acts as a regularization factor.

Finally, the 22 feature vectors extracted by the cascade of convolutional networks (i.e., the features of size $22 \times 12 \times 12 \times 8$) are concatenated and are fed to two fully connected layers ("Fully7" - "Fully8"). ReLU is used as an activation function for these two fully connected layers, also these fully connected layers are mainly used to capture the correlations among these features and compress them into a smaller vector (i.e., 1024-D). After representing the input facial image by a 1024-D features vector, the multi-labels classification is performed by another fully connected layer with c outputs. We use the Sigmoid function as an activation function in this "Output" layer to make each value in the c-D outputs vector representing the prediction $x_j \in [0, \; 1]$ of the j^{th} AU of interest.

This setting of AUs classification is a multi-label learning problem. We choose the following weighted cross-entropy $L(Y, X)$ to be minimized as a loss function. This function measures the probability error in AUs classification.

$$L(Y, X) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} \alpha_j y_{ij} \log x_{ij} + (1 - y_{ij}) \log(1 - x_{ij}),$$

where $X \in \mathbb{R}^{N \times c}$ is the matrix of the output layer responses for N samples. $Y \in \{0, 1\}$ is the matrix of the ground truth labels where each element y_{ij} is the ground truth label of i^{th} sample for j^{th} AU. The weight α_j is multiplied by the first term to up-weight the cost of a positive error relative to a negative error for j^{th} AU. These weights are used to overcome the well-known imbalance data problem i.e., the number of positive samples of AUs is less than the negative ones. Finally, two regularization methods are used to prevent overfitting during the training process: the dropout and the ℓ_2 norm of the weights, which is added to the loss function.

2.1. Patch Significance:

As shown in Fig. 1(c), we choose 22 overlapped patches out of the 49 patches. To show the significance of the selected patches and how these patches affect AUs classification, we apply a method that is similar to the occlusion sensitivity maps approach [30] as

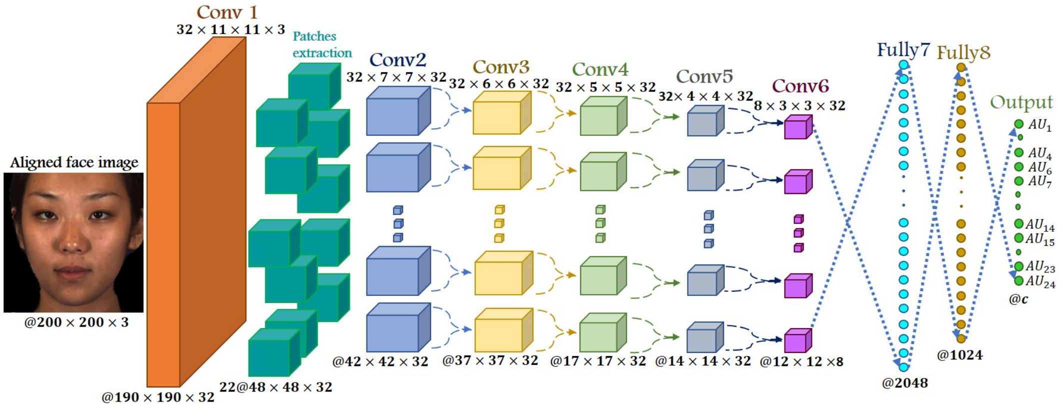


Figure 2. The proposed deep region learning architecture. Low level features are extracted from an aligned RGB facial image by a convolutional layer (Conv1). Then 22 overlapped patches of sizes 48 × 48 are extracted from the convolutional layer output. Each patch is processed by a different cascaded of five convolutional layers (Conv2-Conv6). The filter size of each layer is written on the top and the dimensions of layer's output are written on the bottom. The 22 feature vectors extracted by Conv6 are concatenated and fed to consecutive three fully connected layers to detect c AUs.

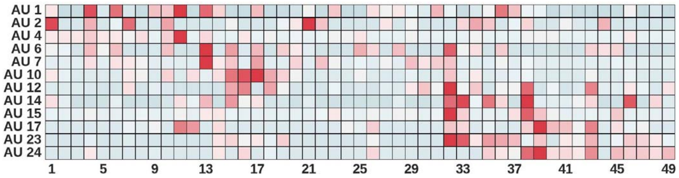


Figure 3. An image illustrates the patches significance (ordered from dark red to dark blue) for each AUs.

follows: the proposed model shown in Fig. 2 is using all 49 patches. Using $30,000$ samples, we calculate the score of each AU, but to occlude a certain patch effect the "Conv6" output of this patch is fed as zeroes to "Fully7". This is sequentially repeated for all 49 patches and the patch significance is calculated as its average effect on the score of a certain AU. The significance of patches for each AU is shown in Fig. 3. Note that the numbers of the patches correspond to the numbers of the landmarks shown in Fig. 1(b). From Fig. 3, we can infer the following facts about the semantic relationships among AUs, which are similar to what have been illustrated in the state-of-the-art e.g., [6], [10]: patches around inner eyebrow 4 and 6 as well as patches in between 11 and 13 are the most significant for "Inner Brow Raiser" AU1; the set of most significant patches for "Outer Brow Raiser" AU2 contains outer brow patches 1 and 10; the high significance of patches 7 for AU2 confirms the positive correlation between AU1 and AU2; and the high significance of patch 32 for "Cheek Raiser" AU6 and "Lip Corner Puller" AU12 highlights the positive correlation between these two AUs. Fig. 3 highlights that lips related AUs, i.e., 12, 14, 15, 23 and 24, have their most significant patches around the lips. These correctly learned correlations among different AUs confirm the effectiveness of the proposed architecture in detecting different action units. The selected 22 overlapped patches include all the regions of interest of these AUs.

3. Experiments

To evaluate the performance of the proposed network, we use two datasets: BP4D-Spontaneous dataset [32] and the recently released FERA17 dataset [31].

BP4D-Spontaneous dataset [32]: This dataset consists of 328 videos that were captured during a series of eight emotional expressions for 23 females and 18 males. The dataset has a frame-based action units coding. We conduct our experiment using videos of 31 subjects as training/validation data and videos of the remaining 10 subjects as testing data. The huge number of frames in these videos are sampled to obtain valid aligned facial images. This sampling reduces the dataset to approximately $110,000$ valid frames.

FERA17 dataset: The range of head movements in the BP4D dataset is moderate. So, recently, the FERA17 dataset [31] was released with 9 different poses. FERA17 has 328 3D sequences for 41 subjects of the BP4D [32]. These are used as a train/validation set. Another 159 3D sequences for 20 subjects that were derived from a subset of BP4D+ database [33] are used as a test-set (This is the development partition of FERA17, which is publicly available). These 3D sequences in BP4D and BP4D+ are rotated by pitch angles ($-40°$, $-20°$, and $0°$) and yaw angles ($-40°$, $0°$, and $40°$). Then nine videos were created, see Fig. 4. Also, the dataset has a frame-based action units coding. We sampled approximately $300,000$ valid frames out of 3896 videos. Videos with poses 1 and 7, shown in Fig. 4, are excluded because the preprocessing step does not generate many valid frames from these videos due to the occlusion in the left eye.

To illustrate the imbalance in these datasets, skew i.e., the ratio of the number of negative samples to the number of positive samples, in these train-sets for each AU is shown in Table 1.

Learning: To train the proposed network, we use an adaptive learning rate optimization method [34]. The initial learning rate is 1.0 and the momentum is 0.95. The weight α_j in the loss function is chosen to be the skew of the corresponding AU in the data. The dropout rate is 50%. The batch size is 128. The weight decay is 0.0005. All experiments were performed on one NVIDIA Titan X GPU.

Performance Measures: We use three metrics as a performance measure: Area under the Precision Recall curve (APR), Area under the ROC curve (AROC), and f_1 score. However, the APR and f_1 score are attenuated by the skewed distributions [25]. Thus, the normalized versions nAPR and nf_1 score of these metrics are calculated.

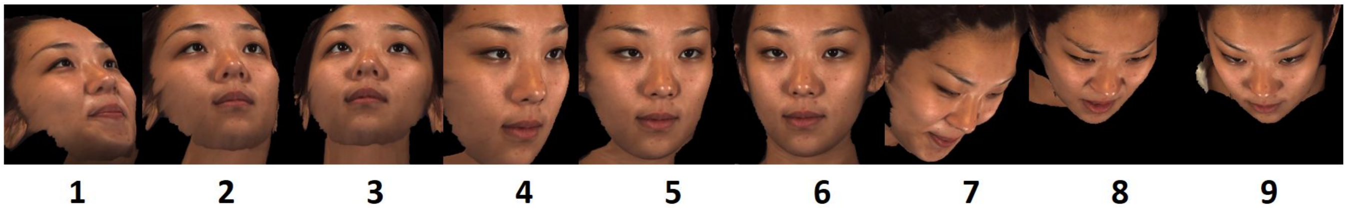


Figure 4. Nine different poses in FERA17 dataset [31].

TABLE 1. SKEW OF DIFFERENT AUS WITHIN BP4D [32] AND FERA17 [31] TRAIN SETS

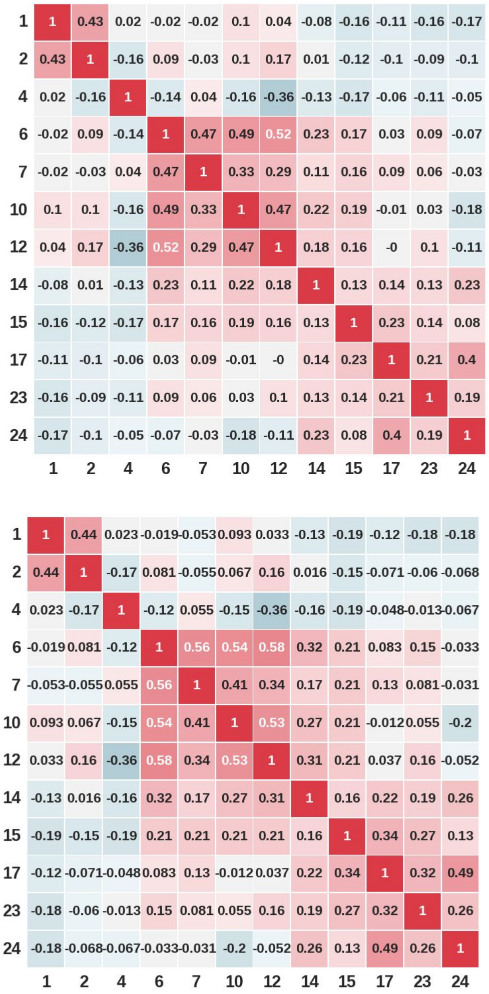
AU	1	2	4	6	7	10	12	14	15	17	23	24
BP4D	2.2	3.4	2.4	1.2	0.8	0.7	0.8	1.0	3.1	1.3	3.4	3.6
FERA17	1.7	-	2.4	1.0	0.6	0.5	0.6	0.9	1.6	0.9	1.4	-

TABLE 2. RESULTS ON THE BP4D TEST-SET USING DIFFERENT PERFORMANCE MEASURES: f_1 SCORE, nf_1 SCORE, APR, NAPR, AND AROC.

AU	1	2	4	6	7	10	12	14	15	17	23	24	Avg
skew	2.4	2.3	2.4	0.9	0.6	0.6	1.0	0.9	3.1	1.3	2.7	3.5	-
f_1	0.51	0.56	0.65	0.78	0.81	0.83	0.81	0.74	0.60	0.68	0.59	0.60	0.68
nf_1	0.68	0.70	0.78	0.77	0.75	0.77	0.81	0.72	0.77	0.72	0.72	0.79	0.75
APR	0.52	0.58	0.72	0.83	0.85	0.85	0.86	0.75	0.61	0.68	0.63	0.51	0.70
nAPR	0.71	0.74	0.84	0.82	0.78	0.76	0.86	0.73	0.82	0.73	0.80	0.77	0.78
AROC	0.69	0.73	0.84	0.84	0.81	0.81	0.87	0.75	0.83	0.75	0.78	0.82	0.79

As a first evaluation, we train the proposed model using the BP4D [32] train-set. To illustrate the capability of the proposed network in learning the semantic relationships among AUs, the relation matrix of the ground truth AUs and the relation matrix of the proposed network predictions are computed. Each relation matrix contains the correlation coefficients between pairwise AUs. These matrices are shown in Fig. 5. The element-wise Euclidean distance 0.004 between the two matrices confirms the ability of the proposed network in learning the semantic relationships of AUs. The trained model is then used to predict the presence of the action units in BP4D test-set. The different performance metrics: f_1 score, nf_1 score, APR, nAPR, and AROC of this experiment are shown in Table 2.

To measure the generalizability of the proposed model in a cross-dataset scenario, we train the proposed model using the FERA17 train-set. This model is then used to predict the presence of the action units in the BP4D test-set. The performance measures are reported in Table 3. The results in Tables 2 and 3 are very close to each other. This confirms that cross-dataset protocol is successfully applied on the proposed model.



	1	2	4	6	7	10	12	14	15	17	23	24
1	1	0.43	0.02	-0.02	-0.02	0.1	0.04	-0.08	-0.16	-0.11	-0.16	-0.17
2	0.43	1	-0.16	0.09	-0.03	0.1	0.17	0.01	-0.12	-0.1	-0.09	-0.1
4	0.02	-0.16	1	-0.14	0.04	-0.16	-0.36	-0.13	-0.17	-0.06	-0.11	-0.05
6	-0.02	0.09	-0.14	1	0.47	0.49	0.52	0.23	0.17	0.03	0.09	-0.07
7	-0.02	-0.03	0.04	0.47	1	0.33	0.29	0.11	0.16	0.09	0.06	-0.03
10	0.1	0.1	-0.16	0.49	0.33	1	0.47	0.22	0.19	-0.01	0.03	-0.18
12	0.04	0.17	-0.36	0.52	0.29	0.47	1	0.18	0.16	-0	0.1	-0.11
14	-0.08	0.01	-0.13	0.23	0.11	0.22	0.18	1	0.13	0.14	0.13	0.23
15	-0.16	-0.12	-0.17	0.17	0.16	0.19	0.16	0.13	1	0.23	0.14	0.08
17	-0.11	-0.1	-0.06	0.03	0.09	-0.01	-0	0.14	0.23	1	0.21	0.4
23	-0.16	-0.09	-0.11	0.09	0.06	0.03	0.1	0.13	0.14	0.21	1	0.19
24	-0.17	-0.1	-0.05	-0.07	-0.03	-0.18	-0.11	0.23	0.08	0.4	0.19	1

	1	2	4	6	7	10	12	14	15	17	23	24
1	1	0.44	0.023	-0.019	-0.053	0.093	0.033	-0.13	-0.19	-0.12	-0.18	-0.18
2	0.44	1	-0.17	0.081	-0.055	0.067	0.16	0.016	-0.15	-0.071	-0.06	-0.068
4	0.023	-0.17	1	-0.12	0.055	-0.15	-0.36	-0.16	-0.19	-0.048	-0.013	-0.067
6	-0.019	0.081	-0.12	1	0.56	0.54	0.58	0.32	0.21	0.083	0.15	-0.033
7	-0.053	-0.055	0.055	0.56	1	0.41	0.34	0.17	0.21	0.13	0.081	-0.031
10	0.093	0.067	-0.15	0.54	0.41	1	0.53	0.27	0.21	-0.012	0.055	-0.2
12	0.033	0.16	-0.36	0.58	0.34	0.53	1	0.31	0.21	0.037	0.16	-0.052
14	-0.13	0.016	-0.16	0.32	0.17	0.27	0.31	1	0.16	0.22	0.19	0.26
15	-0.19	-0.15	-0.19	0.21	0.21	0.21	0.21	0.16	1	0.34	0.27	0.13
17	-0.12	-0.071	-0.048	0.083	0.13	-0.012	0.037	0.22	0.34	1	0.32	0.49
23	-0.18	-0.06	-0.013	0.15	0.081	0.055	0.16	0.19	0.27	0.32	1	0.26
24	-0.18	-0.068	-0.067	-0.033	-0.031	-0.2	-0.052	0.26	0.13	0.49	0.26	1

Figure 5. The ground truth relation matrix of BP4D dataset (top) and the corresponding relation matrix computed by predictions of proposed approach (bottom).

TABLE 3. DIFFERENT PERFORMANCE MEASURES FOR THE PROPOSED MODEL WHEN IS TESTED ON BP4D TEST-SET AND TRAINED ON FERA17 TRAIN-SET

AU	1	4	6	7	10	12	14	15	17	23	Avg
skew	2.4	2.4	0.9	0.6	0.6	1.0	0.9	3.1	1.3	2.7	-
f_1	0.45	0.48	0.64	0.78	0.82	0.83	0.81	0.73	0.48	0.68	0.67
nf_1	0.67	0.67	0.75	0.77	0.77	0.77	0.80	0.70	0.72	0.72	0.73
APR	0.35	0.41	0.71	0.79	0.84	0.87	0.88	0.72	0.44	0.66	0.67
nAPR	0.53	0.60	0.83	0.78	0.76	0.79	0.88	0.69	0.70	0.72	0.73
AROC	0.47	0.60	0.81	0.81	0.80	0.82	0.88	0.72	0.72	0.74	0.74

The other set of experiments are conducted to evaluate the performance of the proposed network under pose variations. The proposed model, which is trained using the FERA17 train-set, is used to predict the presence of the action units in the FERA17 test-set. For each pose, the different performance metrics: nf_1 score, nAPR, and AROC are shown in Fig. 6. As shown in Fig. 7(a), the

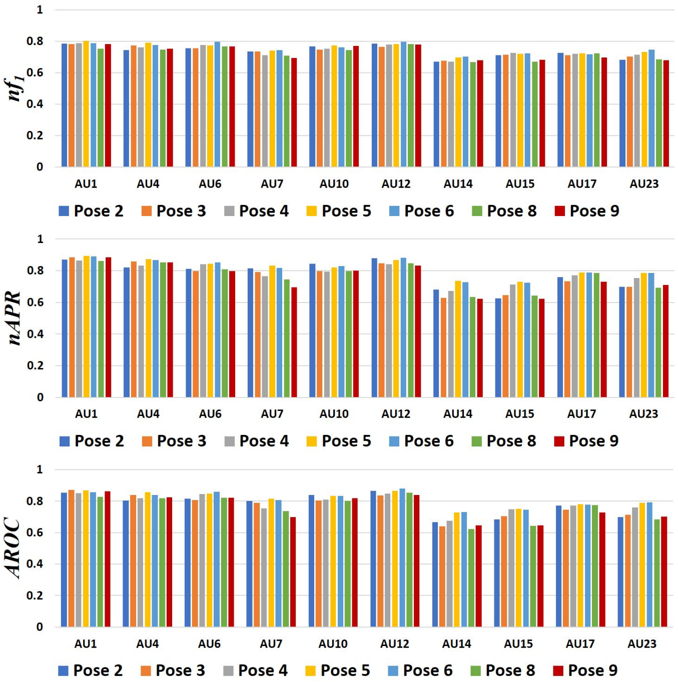


Figure 6. Different performance measures for the proposed AUs detection approach under different poses using FERA17 [31] test-set F1-score (top), Area under PR curve (middle), and Area under ROC curve (bottom).

low standard deviations of the different metrics for the seven poses highlights the pose invariant capability of the proposed model to detect different action units.

Another experiment is conducted to illustrate the significance of the patch-based model. We build a similar model (named "convnet") to the proposed shown in Fig. 2. In this "convnet" model, the region-based layers (i.e.,"Conv2" - "Conv6") are replaced by standard convolutional layers. We keep the sizes of filters as in our model. Similarly, the "convnet" model is trained using the FERA17 train-set with the same settings of our trained model. Then our proposed model and the "convnet" model are used to predict the presence of the action units in the FERA17 test-set. The average over all posses of the different performance metrics: f_1, nf_1 , APR, nAPR, and AROC are shown in Table 4. These performance measures illustrate the following: the "convnet" model has a slight enhancement in only AU6; other action units are more accurately detected using our proposed model than the standard 'convnet". This enhancement is up to 15% in ROC and nAPR of AU23. Moreover, the proposed model has lower standard deviations (see Fig. 7(a)) than the "convnet" model for the different metrics (see Fig. 7(b)). This confirms that our proposed patch-based layer is more effective in capturing the required structural features of the face and learns the correlations among AUs under pose variations than the standard convolutional layer.

Comparison with the-state-of-the-art: The closest work to the proposed one is recently introduced by Zhao et al. [6]. They conducted experiments using the BP4D dataset [32], however, unlike our sampling, they sampled 100 positive and 200 negative frames for each sequence and they adopted 3-fold partition instead of our partitioning. The authors reported (see Table 2 in [6]) the performance of different related work such as the classical linear support vector machine classification, patch-learning method [7], JPML [10] and other deep network-based methods (e.g.,locally connected network, AlexNet [35], and their DRML model [6]). Comparing these reported performance measures to ours (i.e., APR Avg= 70% in Table 2 and DRML Avg= 56% [6]), confirms the high performance of the proposed approach, but using a different setting as explained. Also, the element-wise Euclidean distance 0.004 between the two relation matrices is smaller than what were reported for AlexNet and DRML models in [6]. This confirms that the proposed approach outperforms the state-of-the-art approaches.

TABLE 4. AUs DETECTION PERFORMANCE OF OUR PROPOSED MODEL, THE "CONVNET" MODEL, AND BASELINE RESULTS [31] USING THE FERA17 TEST-SET. DIFFERENT PERFORMANCE MEASURES: APR, nAPR, AROC, f_1 SCORE AND nf_1 SCORE ARE USED.

AU		1	4	6	7	10	12	14	15	17	23
skew		9.6	10.5	1.6	0.5	0.6	1.0	0.4	4.1	2.9	1.9
f_1	convnet	0.50	0.39	0.73	0.82	0.80	0.77	0.82	0.36	0.49	0.54
	FERA17 [31]	0.15	0.17	0.56	0.73	0.69	0.65	0.62	0.15	0.22	0.21
	ours	0.57	0.49	0.70	0.82	0.81	0.77	0.82	0.42	0.54	0.59
nf_1	convnet	0.75	0.73	0.78	0.72	0.74	0.78	0.67	0.67	0.70	0.68
	ours	0.78	0.76	0.77	0.72	0.76	0.78	0.68	0.71	0.71	0.71
APR	convnet	0.51	0.38	0.79	0.87	0.86	0.82	0.78	0.27	0.48	0.48
	ours	0.59	0.50	0.75	0.87	0.87	0.85	0.82	0.34	0.55	0.60
nAPR	convnet	0.85	0.80	0.85	0.78	0.79	0.83	0.62	0.60	0.71	0.63
	ours	0.88	0.85	0.82	0.78	0.81	0.85	0.67	0.68	0.7	0.73
AROC	convnet	0.82	0.79	0.85	0.77	0.79	0.84	0.63	0.62	0.71	0.65
	ours	0.86	0.83	0.83	0.77	0.82	0.85	0.67	0.71	0.76	0.74

It is worth mentioning that Tosér et al. [23] recently conducted a similar experiment for action units detection under pose variations. The authors used an old version of the FERA17 dataset (i.e., FERA15). Our nf_1 scores are better than what have been reported in [23] for the multi-label model. However, since the datasets are different, we cannot consider this as a fair comparison. Finally, as shown in Table 4 action units are more accurately detected using our proposed model than the FERA17 baseline results [31]. The enhancement in the f_1 scores are from 9% to 42%.

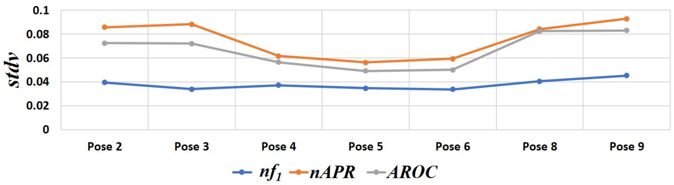
4. Conclusions

We propose a region-based architecture and a multi-label deep learning approach. The proposed approach combines the different state-of-the-art advantages: First, features and AUs correlations are jointly learned using a non-linear model. Second, the proposed work is an end-to-end trainable and multi-label learning approach. Additionally, it overcomes the disadvantage of using a uniform grid. Importantly, by using the weighted loss function, we overcome the skew problem in the AUs data.

The proposed region-based network has the capability of capturing the required structural features of the face and learning the correlations among AUs. This has been verified by many experiments using different datasets. The performance of the proposed approach outperforms the state-of-the-art methods.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU and TSN, Inc. for funding this research.



(a)

(b)

Figure 7. Standard deviations of the different metrics for the 7 poses.

References

[1] C. Darwin, *The Expression of Emotions in Man and Animals*. John Murray 1872, reprinted by University of Chicago Press, 1965.

[2] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[3] C. L. Lisetti and D. J. Schiano, "Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect," *Pragmatics & Cognition*, vol. 8, no. 1, pp. 185–235, 2000.

[4] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement-from facial expressions," *IEEE Trans. Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[5] G. Duchenne, "Mecanisme de la physionomie humaine," *Paris, France: Renouard*, 1862.

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

[7] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybernetics*, vol. 45, no. 8, pp. 1499–1510, 2015.

[8] S. H. Ping Liu, "Facial expression recognition via a boosted deep belief network," in *CVPR*, 2014.

[9] P. Liu, J. T. Zhou, I. W. 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.

[10] 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.

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

[12] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "EAC-Net: A region-based deep enhancing and cropping approach for facial action unit detection," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2017.

[13] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS): Manual*. A Human Face, 2002.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *CVPR-Workshops*, 2010.

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

[16] X. Zhang and M. H. Mahoor, "Task-dependent multi-task multiple kernel learning for facial action unit detection," *Pattern Recogn.*, vol. 51, no. C, pp. 187–196, 2016.

[17] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *ICCV*, 2013.

[18] Y. Tong and Q. Ji, "Learning bayesian networks with qualitative constraints," in *CVPR*, 2008.

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

[20] S. Lucey, I. A. Matthews, C. Hu, Z. Ambadar, F. D. la Torre, and J. F. Cohn, "AAM derived face representations for robust facial action recognition," in *Int. Conf. on Automatic Face & Gesture Recognition*, 2006.

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

[22] S. Jaiswal and M. F. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[23] Z. Tosér, L. A. Jeni, A. Lörincz, and J. F. Cohn, "Deep learning for facial action unit detection under large head poses," in *ECCV-Workshops*, 2016.

[24] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014.

[25] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[26] M. Zhang, Y. Li, and X. Liu, "Towards class-imbalance aware multi-label learning," in *Int. Joint Conf. on Artificial Intelligence*, 2015.

[27] E. Mostafa, A. A. Ali, A. Shalaby, and A. Farag, "A facial features detector integrating holistic facial information and part-based model," in *CVPR-Workshops*, 2015.

[28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification." in *CVPR*, 2015, pp. 2746–2754.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[31] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic, "FERA 2017 - addressing head pose in the third facial expression recognition and analysis challenge," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2017.

[32] X. Zhang, L. Yin, J. F. Cohn, S. J. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database." *Image Vision Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[33] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *CVPR*, June 2016.

[34] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.