Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing

Wei Li
Dept of Electrical Engineering
CUNY City College
New York, NY, USA
wli3@ccny.cuny.edu

Farnaz Abtahi
Dept of Computer Science
CUNY Graduate Center
New York, NY, USA
fabtahi@gradcenter.cuny.edu

Zhigang Zhu
Dept of Computer Science
CUNY City College and Graduate Center
New York, NY, USA
zhu@cs.ccny.cuny.edu

Abstract

Action Unit (AU) detection becomes essential for facial analysis. Many proposed approaches face challenges in three aspects: the alignments of different face regions, the training of a model for multiple AU labels, and the effective fusion of temporal information. To better address these problems, we propose a deep learning framework for AU detection with region of interest (ROI) adaptation, integrated multi-label learning, and optimal LSTM-based temporal fusing. First, ROI cropping nets (ROI Nets) are designed to ensure that specifically interested regions of faces are learned independently; each sub-region has a local convolutional neural network (CNN) - an ROI Net, whose convolutional filters will only be trained for that region. Second, multi-label learning is employed to integrate the outputs of those individual ROI cropping nets, which learns the inter-relationships of various AUs and acquires global features across sub-regions for AU detection. Finally, the optimal selection of multiple LSTM layers to form the best LSTM Net is carried out to best fuse temporal features, in order to make the AU prediction the most accurate. The proposed approach is evaluated on two popular AU detection datasets, BP4D and DISFA, outperforming the state of the art significantly, with an average improvement of around 13% on BP4D and 25% on DISFA, respectively.

1. Introduction

Action Units (AUs) are the basic facial movements that work as the building blocks in formulating multiple facial expressions. The successful detection of AUs will greatly facilitate the analysis of the complicated facial actions or expressions. AU detection has been studied for decades as one of the basic facial computing problems and many interesting approaches have been proposed. Classical approaches in AU detection either focus on facial landmark-based local features or appearance-based global features. A number of deep learning approaches have also been proposed to learn deeper facial representations that result in better AU detection.

However, some essential problems are still not solved completely. Due to different features for different facial components, individual AUs may need to be considered separately. One image may include multiple AUs, therefore whether training single AU or multi-label AUs has to be analyzed. Since all actions appear in a temporal instead of just static mode, fusing temporal information becomes necessary. So, to achieve the best AU detection performance, all the three aspects need to be considered.

Since CNNs have proved to be a powerful tool in solving many image-based tasks and several novel deep structures and frameworks have been proposed, we choose these deep learning models to tackle the AU detection problems. Recently, region-based processing is used in the fast/faster RCNN for prediction of object's bounding box or objectiveness probability in [9, 18]. This inspired us to design separate networks to learn features for different regions of interest on faces. The success in applying LSTM (long and short term memory) in image caption generation [25] and human action recognition [5, 17] led us to believe that it is a good temporal information fusing kernel which may be also useful for facial AU detection.

After identifying the three problems and being inspired by these RCNN and LSTM approaches, we designed an adaptive region cropping based multi-label learning deep recurrent net. The structure of the proposed neural network is shown in Figure 1. There are a number of unique features of the proposed network. Unlike conventional CNNs where the same convolutional filters are shared within the same convolutional layers, we crop individual regions of interest(ROIs) based on facial landmarks from all the feature maps. Each cropped region (as represented by red circle, yellow triangle, gray square and black diamond in the

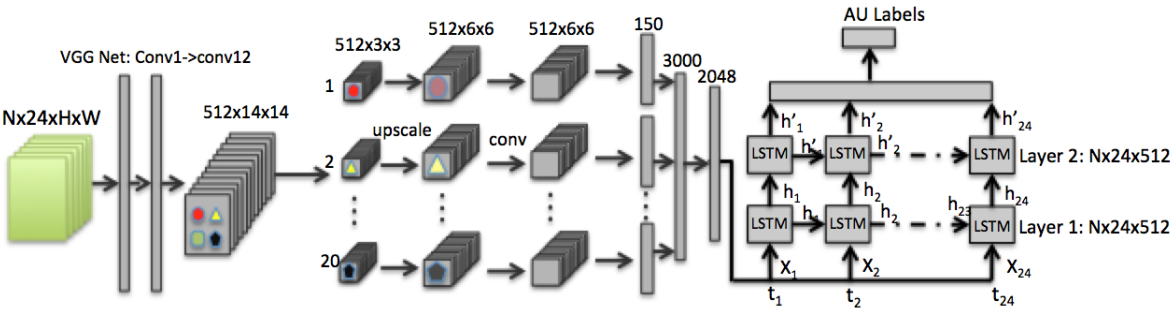

Figure 1. Framework of the proposed neural network with VGG Net, ROI Nets and LSTM Net

figure), represents an area of interest. So, these ROIs are learned individually and therefore important areas will be able to receive special attention. To fuse the temporal information of expressions, the features from the final fully connected layer are fed to several stacks of LSTM layers (two in the figure for illustration purpose only). Then, the temporal features are used to predict all AUs simultaneously. Through this structure, our network can handle both the adaptive region learning and the temporal fusing problems.

Comparing to existing approaches, our approach has the following unique contributions:

1) A set of adaptive ROI cropping nets (ROI Nets) is designed to learn regional features separately. In the proposed network, each ROI has a local convolutional neural network. The convolutional filters will only be trained for corresponding regions.

2)Multi-label learning is employed to integrate the outputs of those individual ROI cropping nets, which learns the inter-relationships of various AUs and acquires global features across sub-regions for AU detection. Multi-label and single AU based methods are compared. With additional AU correlations and richer global features, the multi-label learning approach shows slightly better performance.

3) An LSTM-based temporal fusion recurrent net (LSTM Net) is proposed to fuse static CNN features, which makes the AU predictions more accurate than with static images only.

This paper is organized as the follows. In Section 1, we have introduced the problems in AU detection and the basic idea of our proposed approach. In Section 2, we review the related work on AU detection, including both traditional and deep learning approaches. We then explain our proposed region learning based CNN network in Section 3. Section 4 describes the way the temporal information of the CNN features is fused with the LSTM layers. Experimental results are included in Section 5 where we evaluate our proposed approach in terms of regions cropping, multi-label learning and temporal fusion, and performance comparison against baseline approaches are also given. We conclude the paper in Section 6.

2. Related Work

AU detection has been studied for decades and various approaches have been proposed for this problem. Facial key points (landmark points) play an important role in AU detection. Two types of features were usually used in landmark-based approaches. Landmark geometry features were obtained by measuring the normalized facial landmark distances and the angles of the Delaunay mask formed by the landmark points. On the other hand, landmark texture features were obtained by applying multiple orientation Gabor filters to the original images. Many conventional approaches [6, 14, 26, 2, 4, 28, 16, 24] were designed by employing texture features near the facial key points. Valstar et al [23] analyzed Gabor wavelet features near 20 facial landmark points. The features were then selected and classified by Adaboost and SVM classifiers. Since landmark geometry has been found robust in many AU detection methods, Fabian et al [1] proposed an approach for fusing the geometry and local texture information. Zhao et al [29] proposed the Joint Patch and Multi-label Learning (JPML) method for AU detection. Similarly, landmark-based regions were selected and SIFT features were used to represent the local patch. Overall, the conventional approaches focused on designing artificial features near facial areas of interest. The appearance changes, representing the motion of the landmark points, give an indication of the facial action units.

In addition to facial AU detection, some researchers have also focused on other related problems. Song et al [20] investigated the sparsity and co-occurrence of action units. Wu et al [27] explored the joint of action unit detection and facial landmark localization and showed that the constraints can improve both AU and landmark detection. Girard et al [8] analyzed the effect of different sizes of training datasets on appearance and shape-based AU detection. Gehrig et al [7] tried to estimate action unit intensities by employing linear partial least squares to regress intensities in AU related regions.

Over the last few years, we have witnessed that CNNs boost the performance in many computer vision tasks. Compared to most conventional artificially designed features, CNNs can learn and reveal deeper information from training images. Deep learning has also been employed for AU detection [15]. Two pieces of the most recent work on the use of deep learning for AU detection are noteworthy. Zhao et al [30] used a deep learning approach by dividing aligned face images into 8x8 blocks. These 64 separate areas are then learned separately. However, although this approach worked well for each individual part of a face, it highly relied on face alignment. Additionally, treating all blocks equally may degrade the importance of some regions. Chu et al [3] proposed a hybrid approach for combining CNN and LSTM to learn a better representation of an AU sequence. Due to the fusion of both spatial CNN and temporal features, the AU detection performance in this work has improved significantly compared to existing approaches. However, the proposed network is a conventional CNN, which is unable to extract local features from specific regions. Jaiswal et al [12] proposed a dynamic appearance and shape based deep learning approach. A shallow region and shape mask CNN is employed to learn the static feature while LSTM is used to extract a dynamic feature from the trained CNN model. In our work, we have designed a CNN which can not only focus on different facial regions independently but also fused the temporal features using recurrent networks.

3. Region of Interest Learning: ROI Nets

CNNs have recently been the most popular tool for image understanding. In a classic CNN structure, a convolutional layer is composed of multiple filters and activation functions. The convolutional filters cover the entire image and generate corresponding feature maps. In this manner, convolutional filters are shared by all the regions of the feature maps. This approach is effective in dealing with general image feature detection, but for some tasks in which individual local regions should be treated differently, sharing the same set of filters for the entire image is not an effective approach. As most traditional approaches tried to find local SIFT or Gabor features near facial landmark points, we would like to learn local CNN features in these regions of interest (ROIs).

We use the BP4D dataset for AU detection which includes 12 AUs. The index, name and corresponding muscles of each AU are illustrated in Table 1 for all the 12 AUs. The corresponding 2D positions of these AUs are shown in Figure 2. We first use a landmark detection algorithm [13] to find the facial landmark points on a face (blue points in Figure 2 right). We choose the AU centers based on the positions of the related muscles (Figure 2 left), which are adjusted from face to face using the detected facial landmark points. Note that some landmark points are not in the centers of facial action muscle regions but they are close to them and can be used to locate the muscles. In the end, the center of an AU is either at a landmark point or a certain distance away from a landmark point, as shown with a pair of blue-to-green point in the figure; we used 20 landmark points in total.

Table 1. Rules for defining AU centers

AU index	Au Name	Muscle name
1	Inner Brow Raiser	Frontalis
2	Outer Brow Raiser	Frontalis
4	Brow Lowerer	Corrugator supercilii
6	Cheek Raiser	Orbicularis oculi
7	Lid Tightener	Orbicularis oculi
10	Upper Lip Raiser	Levator labii superioris
12	Lip Corner Puller	Zygomaticus major
14	Dimpler	Buccinator
15	Lip Corner Depressor	Triangularis
17	Chin Raiser	Mentalis
23	Lip Tightener	Orbicularis oris
24	Lip Pressor	Orbicularis oris

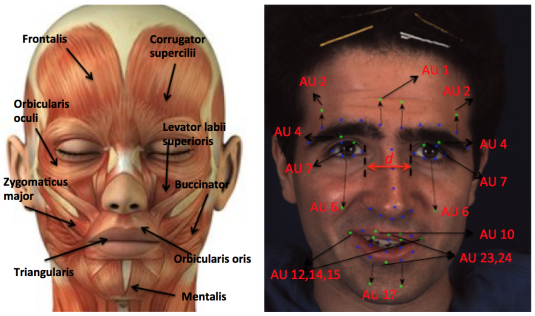

Figure 2. ROI center selection based on muscles and landmarks

Knowing the landmark positions, we can then design the neural network cropping layers to form the ROI Nets. We use VGG [19] as the base for our ROI Net due to its simple structure and excellent performance in object classification. We also choose the 12th convolutional layer as the feature map for cropping. We finally crop the face into 20 ROIs for separate AU learning. In other words, the 20 green points in Figure 2 are regarded as ROI AU centers.

The corresponding positions of AU centers in the feature map can be found based on the ratio of the original image size (224×224) and the feature map size (14×14). Based on the $512\times14\times14$ feature maps as well as the 20 AU centers, we take a total of 20 sub-regions (centered at the selected AU landmark centers), each of $512\times3\times3$, as the input for cropping layers to form the ROI-Nets, 20 in total. For each individual region learning network, the input size of 3×3 might not be able to represent the region well: If the 3x3

feature map is directly connected to a 3×3 filter, it will be turned into a single value, hence spatial information is lost. So, upsampling layers are added to upscale these $3times3$ feature maps to 6×6 before the convolutional layers. As a result, applying the 3x3 filter to an upsampled 6×6 feature map yields a 4×4 feature array (This upsampling actually leads to a 1% improvement of the average F1 score). The final adaptive region learning structure is shown in the middle part of Figure 1. After local learning with the ROI Nets, we use a fully connected feature vector to represent the local regional features. Then we can either pair the symmetrical features for single AU detection or concatenate all the fully connected features for multi-label AU detection. We will conduct a further comparison on this selection in Section 5.

By designing the ROI Nets, we can train separate filters for the AUs. This may make the feature learning adaptive to different local facial properties. Comparison of the ROI learning and conventional CNN learning will be performed in the evaluation section (Section 5).

4. Temporal Fusing: LSTM Net

A facial action always has a temporal component when using a video sequence as the input, hence knowing the previous states of a facial expression can definitely improve the AU detection. However, one of the limitations of the CNN structure is the lack memory of previous states. Regular CNNs are only able to process a single image at a time. To deal with a sequence of images, C3D [22], which is basically a 3D version of CNN, has been proposed. C3D can deal with sequential images but the number of input images is fixed. The training of a C3D is very time consuming too. Another huge shortage of C3D is, compared to using regular CNN, the lack of existing pretrained models similar to VGG [19], GoogleLeNet [21] and ResNet [11], which can all provide very good initial parameters as a starting point for training. The current best network for temporal fusion is the Long Short Term Memory (LSTM) network [10]. As a recurrent net, it can memorize the previous features and states, which can help current feature learning and estimation. It also has gate structures to make it suitable for long time and short time temporal feature learning. LSTM has also proved to be effective in action recognition [17].

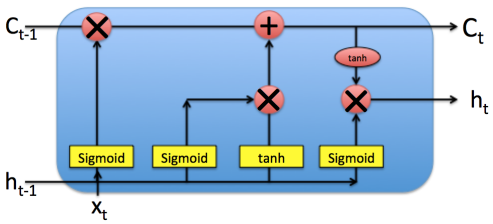

Figure 3. Structure of a simple LSTM block.

The structure of an LSTM block is shown in Figure 3. In the LSTM block, C_{t-1} and C_t are the cell parameters at the previous and the current times, the long and short memories are described by the cell state vector C_t. The cell states store the memory parameters in LSTM. At each time step, an LSTM kernel will take the previous output h_{t-1} and the new input x_t to generate the new output h_t based on kernel parameter C_t. Meanwhile, the cell state C_t gets updated by dropping old information and getting new information. A new input feature fed to a LSTM block will go through three steps: C_t forgets; C_t updates; h_t updates. First, the LSTM has to decide what information to keep/forget from the old cell state. This is based on the previous LSTM output h_{t-1} and new input feature x_t. The forget vector f_t follows equation 1:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

where W_f and b_f are the forget gate parameters. The next step is to update the cell state with new information for future use. The new cell state C_t is determined by three elements: previous partially saved cell state C_{t-1}, current LSTM input x_t and previous output h_{t-1}. The last two vectors need to go through an "input gate" and a $tanh$ activation function. The updated cell state can be obtained using equation 2:

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \tag{2}$$

where i_t is the merged input of x_t and h_{t-1} defined by

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

where W_i and b_i are the input gate parameters. $\check{C}_t$ in equation 2 is the candidate cell state for generating final cell state and output, which can be regarded as a temporal cell state parameter, following equation 4:

$$\check{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

where W_c and b_c are the candidate gate parameters.

Finally, we generate the current output h_t for the LSTM based on the updated cell state C_t, the current input feature x_t and the previous output h_{t-1}, which can be described by

$$h_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \cdot tanh(C_t) \tag{5}$$

where W_o and b_o are the output gate parameters. Meanwhile, the output h_t and the cell C_t are passed to next time output generation.

LSTM can be easily connected to the CNN structure. Fully connected layers of a CNN can be directly fed into the input of LSTM blocks. To better represent the fully connected features, multiple LSTM kernels can act as a layer to represent temporal features. As shown in Figure 1, the CNN model turns each image image into a feature 1-D vector. The first frame of an image sequence at time t_1 is sent to

the LSTM layer at t_1. The LSTM layer will produce output feature h_1 for the first frame, then at time t_2, a new frame is sent to the LSTM layer and the new output feature is produced based on x_2 and h_1, and so on so forth. Here we use $h_i(i = 1...n)$ to represent the ith LSTM feature; in Figure 1 n = 24 (the number of frames in a temporal sequence). In different tasks, either only the last LSTM feature h_n or the whole LSTM features $\{h_1, h_2, ...h_n\}$ are used for final prediction. In our case, we believe that all the frames can contribute to the AU detection. Therefore, we use all the LSTM features; in our experiments, the number of frames is 24.

LSTM can effectively fuse the temporal information in a sequence. Similar to the convolutional layers, more than one LSTM layers can be stacked to form an LSTM Net in order to achieve a deeper understanding of the temporal relationships. As shown in Figure 1, the LSTM Net has 2 LSTM layers stacked for AU detection. To see if LSTM is useful in AU detection, we have conducted experiments to compare LSTM-based temporal fusion versus static image AU prediction. In order to find the best structure of LSTM, we also compared different depth of LSTM layers in Section 5 below.

5. Experimental Evaluation

5.1. Datasets and Metrics

Dataset. AU datasets are harder to obtain compared to other tasks such as image classification. The reason is that there are multiple AUs on one face which requires much more manual labeling work. Here we give a brief review of the AU datasets referred by and compared in this paper.

(1) DISFA: 26 human subjects are involved in the DISFA dataset. The subjects are asked to watch videos while spontaneous facial expressions are obtained. The AUs are labeled with intensities from 0 to 5. We can obtain more than 100,000 AU-labeled images from the videos, but there are much more inactive images than the active ones. The diversity of people also makes it hard to train a robust model.

(2) BP4D: There are 23 female and 18 male young adults involved in the BP4D dataset. Both 2D and 3D videos are captured while the subjects show different facial expressions. Each subject participates in 8 sessions of experiments, so there are 328 videos captured in total. AUs are labeled by watching the videos. The number of valid AU frames in each video varies from several hundred to thousands. There are around 140,000 images with AU labels that we could use.

To train a deep learning model, we need a larger number of image samples, and the diversity of the samples is also important. Following a common experimental setting in the AU detection community, we choose BP4D to train our model and conduct a 3-fold cross validation. We first split the dataset into 3 folds based on subject IDs. Each time two folds are used for training and the third fold for testing. For the DISFA dataset, we use the trained model from BP4D to directly extract the last fully connected layer feature with a length of 2048 to represent the images in DISFA. We run the same cross-validation evaluation experiments as the ones we performed with BP4D based on the extracted features using BP4D.

Metrics. One part of our task is to detect if the AUs are active or not, which is a multi-label binary classification problem. For a binary classification task especially when samples are not balanced, F1 score can better describe the performance of the algorithm [24, 6]. In our evaluation, we compute F1 scores for 12 AUs in BP4D and 8 AUs in DISFA. F1 scores can be compared directly as an indicator of the performance of different algorithms on each AU. The overall performance of the algorithm is described by the average F1 score.

5.2. Adaptive Learning vs. Conventional CNN

We proposed our ROI Nets for the adaptive region learning in Section 3. Compared to the conventional CNNs which share the same set of convolutional filters for the whole feature map, we hypothesize that by learning ROIs separately, a better understanding of AUs can be achieved. To validate this hypothesis, we train 2 neural networks on the BP4D dataset: a fine-tuned VGG model - FVGG, and the ROI Nets (on top of the basic VGG model). 12 AUs are used together, so the loss function is based on the predicted results for the 12 AUs. To prevent extreme loss explode which will stop the training, we added offsets to the loss function as

$$Loss = -\Sigma(l \cdot \log(\frac{p + 0.05}{1.05}) + (1-l) \cdot \log(\frac{1.05 - p}{1.05})) \quad (6)$$

where l is the label and p is the generated probability for an AU.

The two models are both based on static images. During each iteration, we randomly select 50 images as a batch to compute the training loss. SGD is employed for back propagation. The VGG net pretrained parameters are used for initializing the model, and the parameters of the first 8 convolutional layers are not updated during training. This makes the set of parameters smaller, which helps the training algorithm converge. We use the proposed structure (VGG Net + ROI Nets) in Section 3 to train the adaptive region learning mode – which is still called ROI Nets. The new designed regional convolutional filters are initialized following a Gaussian distribution. For the conventional fine-tuned VGG (FVGG) net, only the last prediction layer of the basic VGG model is replaced with a fully-connected layer with 12 kernels. We use sigmoid activation functions for the 12

AU probability generators. The two deep models both start with the same learning rate 0.001 which is decreased when the loss is stable. Momentum for both models is set to 0.9.

The final models with both the ROI Nets and FVGG are obtained after training the deep net 20,000 times. We then compare the F1 scores for each AU. The results are shown in Figure 4. We can see that region learning with ROI Nets yields significant improvement over FVGG, on average by 12.4%.

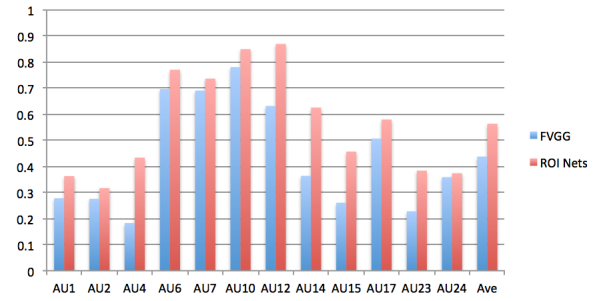

Figure 4. Comparison of FVGG and ROI-Nets in AU detection on BP4D

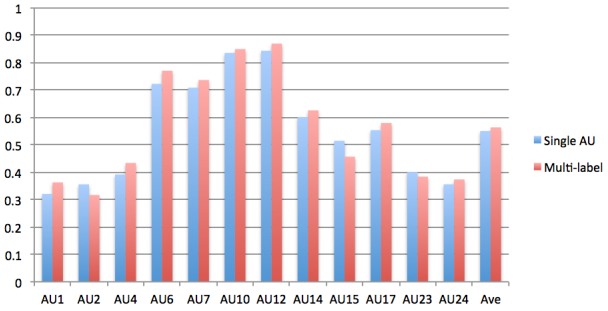

Figure 5. Comparison of single and multi-label learning on BP4D

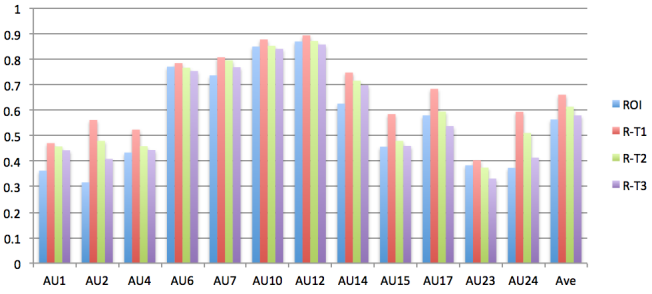

Figure 6. Comparison of static image and temporal fusion in AU detection on BP4D

5.3. Single vs. Multi-label AU Detection

In our proposed ROI Nets, the regions are determined based on the positions where the AUs take place. Since each AU has corresponding regions, we may use only the local learned features to represent the AU for detection. This single AU detection approach differs from the approach we use for the adaptive region learning evaluation (Figure 1) where we concatenate all the AUs features as one fused feature. Our hypothesis is, by concatenating multiple AU features, we may obtain valuable global information as a supplement for individual AU detection or to provide more correlations. However, it's also possible that it brings some noise to the "purity" of an AU feature. To validate our hypothesis, we conduct an experiment to compare single AU detection and multi-label AU detection. In multi-label AU detection, one image is labeled with multiple AUs. In this case, we cannot guarantee that we are able to provide the same number of positive and negative samples for all AUs. But for single AU detection, since the training for each AU is performed separately, we can prepare the training data for each AU in a way that the training data is always balanced during training. The AU detection results for single vs multiple AU detection is shown in Figure 5.

By comparison, we can clearly see that even without equal positive and negative sample distributions, the multi-label AU detection slightly outperforms the single AU detection approach in most AUs, on average by 1.3%. That implies that the global information does have an impact on the fusion learning. We have some more interesting findings if we look into the different AU detection results. For the under-represented AUs (where the AU shows up less frequently in the dataset), such as AU2, AU15, AU23, the balancing of training samples (as in the single AU detection) can boost the performance more significantly. Whereas for some highly related AUs such as AU6 and AU12, both for happy, the multi-label learning has a higher chance to learn this correlation and improve the AU detection for these two AUs.

5.4. Temporal vs. Static

A facial action always has a temporal component, hence knowing the previous state of a facial expression can definitely improve the AU detection. We proposed the LSTM layers for fusing the temporal information with static image features; 512 LSTM kernels are employed to construct each LSTM layer. From our previous evaluations, the best performance was obtained for static images with the ROI Nets. In this experiment, we use the ROI model as a baseline to compare with region cropping recurrent temporal model.

The number of frames fed to the LSTM layers is set to 24 as we follow the settings in most LSTM based action recognition approaches (usually 15-30). More frames may produce better results, but this requires more computing resources. In terms of selecting the 24 frames, We actually have tried using the preceding 23 frames (plus the current frame), but it turned out that the final results were similar to just using one target frame probably due to the very close

Table 2. F1 score on BP4D dataset (ROI: ROI Nets; R-Ti: ROI Nets + i-layer LSTM Net)

AU	LSVM	JPML[29]	DRML[30]	CPM[28]	CNN+LSTM[3]	FVGG	ROI	R-T1	R-T2	FERA[12]
1	23.2	32.6	36.4	43.4	31.4	27.8	36.2	**47.1**	45.8	*28*
2	22.8	25.6	41.8	40.7	31.1	27.6	31.6	**56.2**	48.0	*28*
4	23.1	37.4	43.0	43.4	**71.4**	18.3	43.4	52.4	45.9	*34*
6	27.2	42.3	55.0	59.2	63.3	69.7	77.1	**78.5**	76.7	*70*
7	47.1	50.5	67.0	61.3	77.1	69.1	73.7	**80.8**	79.6	*78*
10	77.2	72.2	66.3	62.1	45.0	78.1	85.0	**87.8**	85.3	*81*
12	63.7	74.1	65.8	68.5	82.6	63.2	87.0	**89.4**	87.2	*78*
14	64.3	65.7	54.1	52.5	72.9	36.4	62.6	**74.8**	71.6	*75*
15	18.4	38.1	36.7	34.0	33.2	26.1	45.7	**58.5**	48.0	*20*
17	33.0	40.0	48.0	54.3	53.9	50.7	58.0	**68.4**	59.5	*36*
23	19.4	30.4	31.7	39.5	38.6	22.8	38.3	**40.4**	37.5	*41*
24	20.7	42.3	30.0	37.8	37.0	35.9	37.4	**59.4**	51.1	-
Avg	35.3	45.9	48.3	50.0	53.2	43.8	56.4	**66.1**	61.4	*51.7*

similarity of nearby frames (within 1 second). So our solution here is to randomly obtain additional 23 frames temporally before to the current frame. The random selection of only 23 samples made them more representative for the videos, more effective in computing and also can provide a larger number of non-redundant training data for deep learning. The LSTM Net is trained after the ROI Nets are trained, which means that we first obtain the ROI model and then the ROI based features are used to train the LSTM model. We do this to make it easier for the model to converge; jointly training the CNN & LSTM might be more effective and this is an interesting future work.

To find the best LSTM structure, we tried 1 (in R-T1), 2 (in R-T2) and 3 (in R-T3) stacked LSTM layers for AU detection, as demonstrated in Figure 1. The AU detection results are shown in Figure 6. We can clearly observe the improvement in AU detection due to applying the LSTM layers. By comparison, R-T1 gives the best performance: the average F1 score is also improved by 9.7% using R-T1 over ROI Nets. Another conclusion we can make here is that with more LSTM layers, the performance decreases, as the ROI features are sufficient to represent the AU images and one LSTM layer is enough to reveal the temporal corrections.

5.5. Comparison with Baselines

To compare our approaches with state of the art methods, we have collected the F1 measures of the most popular methods in same 3-fold settings based on BP4D (Table 2). The approaches includes a traditional SVM-based method, a 2-D landmark feature based approach, JPML [29], the Confidence Preserving Machine (CPM) [28], DRML – a block-based region learning static CNN [30], and CNN+LSTM – a recurrent net fusing LSTM with simple CNN [3].

Table 3. F1 score on DISFA dataset

AU	LSVM	APL[30]	DRML[30]	FVGG	ROI	R-T1
1	10.8	11.4	17.3	32.5	41.5	**42.6**
2	10.0	12.0	17.7	24.3	26.4	**27.2**
4	21.8	30.1	37.4	61.0	**66.4**	65.5
6	15.7	12.4	29.0	34.2	50.7	**55.5**
9	11.5	10.1	10.7	1.67	8.5	**22.8**
12	70.4	65.9	37.7	72.1	**89.3**	82.9
25	12.0	21.4	38.5	87.3	**88.9**	88.3
26	22.1	26.9	20.1	7.1	15.6	**25.9**
Avg	21.8	23.8	26.7	40.2	48.5	**51.3**

For our proposed approaches, we first use the FVGG as the baseline approach. Then, we show the results of adaptive ROI Nets based on static images. Finally, we test our ROI Nets + our LSTM based recurrent approach with one and two LSTM layers (RC+T1, RC+T2). All the results can be seen in Table 2. On average, our best model R-T1 achieves a 12.9% improvement compared to the state of the art approaches. Across the 12 AUs, our R-T1 model outperforms the best in the literature except for AU4, where CNN+LSTM performs the best.

We have also compared with the FERA 2015 BP4D-based challenge winner approach (the last column in Table 2) [12]. Our approach shares the same setting with its baseline using the same training and developing data whereas the winner result was based on test dataset; but we hope this could show the potential of our approach: it outperforms both its baseline (not shown here) and the winner's result with most of the AUs.

To further explore the capabilities of our proposed approach, we run the comparison on DISFA dataset as well. Not as popular as BP4D, fewer state of the art approaches have reported their results on DISFA. Based on our best knowledge, there are less human subjects involved in

DISFA and the average AU occurrence rate is smaller than that of BP4D, which is insufficient to gain good training results. Therefore we use the BP4D trained model (as in a state of the art approach [30]) to extract features from all the images in DISFA and conduct a 3-fold cross evaluation with the extracted features. For static image evaluation, we directly run a multi-label linear regression and for temporal evaluation, we use the structure that shows the best performance in the BP4D evaluation, that is, a one layer LSTM to train the DISFA temporal model. The results are shown in Table 3. As we can see, our R-T1 model leads to a 25% improvement over the state of the art model.

5.6. Discussions

From the results in Tables 2 and 3, our proposed approaches have the best performance in both static and sequence image based AU detection. In the static images based AU detection using deep learning, our ROI Nets outperforms the state of the art deep learning approach, DRML. Our proposed adaptive region cropping method shares the same idea of learning different sets of convolutional filters for different sub-regions, but our method has the following advantages that make it different from the state of the art:

1) Our sub-region selection is adaptive. DRML used a straightforward image dividing strategy. Assuming the facial images are aligned, each image is equally divided into $8times$8=64 sub-regions. This framework in easy to implement, but we have to make sure that the face images are actually aligned in the first place. In order to assure this precondition, all the faces need to be transformed to a neutral shape. This may cause information loss since the faces of different individuals may have different shapes or sizes. In addition, if the original faces are not in a frontal pose, we may also lose some appearance features after changing the pose. On the contrary, we select the regions of interest adaptively. Our approach works based on the detected landmarks and the positions of facial action muscles, which are biologically meaningful. Also note that our approach is robust to landmark position errors. This is because the feature maps in our network go through several pooling layers. Imagine that the position detection error in the original image of size 224×224 is 10 pixel. With the pooling layer for cropping the feature map being of size 14×14, the error turns to be less than 1 pixel. This significantly improves our proposed adaptive region cropping net.

2) Our ROI Nets use learning transfer. A very deep pretrained network (VGG) is used as the base. DRML creates a shallow convolutional network for the region based AU detection. Instead of training everything from scratch, we choose to borrow parameters from an existing very deep CNN model. The main advantage of this approach is that the pretrained model has been trained with millions of images. Although the tasks are different, the parameters are transferable. With the pretrained model as the starting point of our AU detection training, we can achieve a more powerful model than by training a shallow neural network.

In sequential image based AU detection, Chu et al [22] designed a network by combining both CNN and LSTM. To obtain the spatiotemporal fusion features, the last layer features of the CNN and LSTM nets are concatenated. Similarly, Jaiswal et al [12] proposed using a CNN with a shallow region and shape mask to learn static CNN features while LSTM is used to extract dynamic features from the trained CNN model. Different from their uses of AlexNet and a simple CNN for static image feature extraction, we have proposed the adaptive region cropping convolutional nets on top of a more sophisticated CNN model: VGG. We have also used LSTM to fuse the temporal deep features as well, but we have also compared different layers of LSTM and observed that one layer LSTM shows the best performance based on experiments.

6. Conclusion

In this paper, we have investigated three essential problems in AU detection: region adaption learning, temporal fusion and single/multi-label AU learning. We have proposed a novel approach to address these problems: We first proposed the adaptive region of interest cropping nets, which compared to conventional CNN, has been proven to be able to learn separate filters for different regions and can improve the accuracy of AU detection. We then analyzed the proposed model by training it in a multi-label AU detection manner and showed that the new model can outperform a single AU detection model. We finally explored the LSTM-based temporal fusion approach, which boosted the AU detection performance significantly, compared to static image-based approaches. We also tried to find an optimal structure of LSTM layers to connect with the proposed ROI nets to achieve the best results for AU detection. The proposed approach is evaluated on two popular AU detection datasets: BP4D and DISFA, outperforming the state of the art significantly, with an average improvement of around 13% and 25% on BP4D and DISFA respectively. Our future work will be focused on building a dataset-independent AU detection model and applying it to facial action detection in real world applications.

7. Acknowledgement

This work is supported by the National Science Foundation through Award EFRI -1137172, and VentureWell (formerly NCIIA) through Award 10087-12.

References

[1] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16), Las Vegas, NV, USA*, 2016. 2

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013. 2

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016. 3, 7

[4] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2400–2407, 2013. 2

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015. 1

[6] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3792–3800, 2015. 2, 5

[7] T. Gehrig, Z. Al-Halah, H. K. Ekenel, and R. Stiefelhagen. Action unit intensity estimation using hierarchical partial least squares. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–6. IEEE, 2015. 2

[8] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre. How much training data for facial action unit detection? In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015. 2

[9] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1

[10] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. 4

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[12] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016. 3, 7, 8

[13] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 3

[14] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010. 2

[15] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv preprint arXiv:1702.02925*, 2017. 3

[16] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013. 2

[17] S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1942–1950, 2016. 1, 4

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4

[20] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015. 2

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 4

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015. 4, 8

[23] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE, 2006. 2

[24] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015. 2, 5

[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 1

[26] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013. 2

[27] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016. 2

[28] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3622–3630, 2015. 2, 7

[29] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015. 2, 7

[30] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 3, 7, 8