Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 2, 2023, 213-220.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Improved Facial Action Unit Recognition using Local and Global Face Features

Amin Rahmati Sardashti and Foad Ghaderi*
Human-Computer Interaction Lab., Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.

Article Info

Article History:
Received 17 July 2022
Revised 16 October 2022
Accepted 10 January 2023

DOI:10.22044/jadm.2023.12122.2364

Keywords:
Facial action recognition, Facial action units, Deep learning.

**Corresponding author: fghaderi@modares.ac.ir (F. Ghaderi).*

Abstract

Every facial expression involves one or more facial action units appearing on the face. Therefore, action unit recognition is commonly used to enhance the facial expression detection performance. It is important to identify subtle changes in the face when particular action units occur. In this paper, we propose an architecture that employs the local features extracted from specific regions of face while using the global features taken from the whole face. To this end, we combine the SPPNet and FPN modules to architect an end-to-end network for facial action unit recognition. First, different predefined regions of face are detected, and next, the SPPNet module capture deformations in the detected regions. The SPPNet module focuses on each region separately and cannot take into account possible changes in the other areas of the face. In parallel, the FPN module finds the global features related to each of the facial regions. By combining the two modules, the proposed architecture is able to capture both the local and global facial features, and enhance the performance of action unit recognition task. The experimental results on the DISFA dataset demonstrate the effectiveness of our method.

1. Introduction

People's intentions, expressions, physical or mental states usually appear in their faces, and it is believed that people's face say a lot about them. Facial behavior analysis is one of the most popular research areas in affective computing, human-computer interaction (HCI), and machine vision. Previous research works show that people can not completely prevent their intentions and internal states from being represented on their faces [1]. That means, analyzing people's facial behavior can help us understand their goals and intentions.

As mentioned in [2], facial behaviors can be described using two different approaches, i.e., facial expressions and facial action units (AU). Facial expressions are nothing except occurrence of meaningful combinations of facial AUs, which are movements of one or more facial muscles. combinations of two or more different AUs and their appearances on the face depict unique facial expressions. Mapping between the combination of AUs and the corresponding facial expressions is presented in [3]. Using such a mapping, if the AUs and their combinations are identified in a face, facial expressions can also be practically distinguished.

In an attempt for systematic analysis of human facial behavior, Ekman and Friesen developed the facial action coding system (FACS), which is a comprehensive reference system for studying facial actions based on anatomy of human face [4]. The goal of AU detection in a given facial image (or in a sequence of frames in a video) is to measure the similarity of facial muscle movements with those defined in FACS.

Action unit detection is a difficult task and no one can perform it with high performance if they don't have prior knowledge. Nevertheless, manual annotation is time-consuming and expensive, such that it takes more than 30 minutes for an expert to

annotate one minute of a video clip [5]. Moreover, subtle changes in parts of face during the AU occurrence yield to variations in AU appearance, which causes more challenges in the AU recognition task. On the other hand, there are more technical challenges in automatic AU detection, namely, lack of large datasets with AU annotations, diverse subjects, and imbalanced AU datasets.

The action unit recognition methods can be divided into two group, i.e., those that use the whole face, and those that first divide the face image into parts related to the AUs and afterwards classify each part separately. In the latter, it is possible to tackle the subtle facial variations more carefully, however, global features and the relations and the dependencies among the AUs are missing. In this work, we combined the two approaches to take advantage of both local and global information simultaneously. This can help to obtain higher recognition rates and eliminate possible flaws.

The rest of the paper is organized as what follow. An overview of the previous research works in the field of facial AU detection is listed in Section 2. Our proposed method is presented in detail in Section 3. The experimental results are discussed in Section 4. Finally, the paper is concluded in Section 5.

2. Related Works

Many efforts have been made over the previous years in the research filed of AU detection to extract useful features for enhancing detection rate. Static two-dimensional image representation is one of the famous methods of facial AU feature extraction [6]. In this approach, facial features are divide into two categories i.e., appearance and geometry. Gabor wavelets [7, 8], Haar feature [9], scale-invariant feature transform (SIFT) [10], and local binary pattern [11] are the most common handcrafted appearance-based features. On the other hand, deformations in the various components of face convey information that constitute geometric features and can be measured by optical flows [12] or dislocation of landmark points [13, 14]. Some researchers have used a combination of these two feature representation approaches to improve the overall performance [6]. The authors of [15] proposed Multiple kernel Learning. The authors of [16] used the SimpleMKL algorithm, combined the two types of features, and averaged the outcome to exploit the temporal information in sequences. The authors of [17] proposed a multi-conditional latent variable model that encodes the AUs dependencies at both feature and model level into the proposed manifold learning for AU recognition by introducing topological and relational constraints.

The power of deep learning algorithms and their efficiency in various fields has led to the recent use of these techniques in the AU recognition task. The authors of [18] proposed AU R-CNN, in which by designing the AU partition rule, the images are decomposed into a bunch of AU-related bounding boxes and different regions of face are localized. The regions are then merged to obtain the image-level prediction. The authors of [19] proposed a hybrid CNN-RNN network for human action recognition from video. Shao et al. suggested to jointly perform AU recognition and face alignment in order to use the specific AU positions provided by landmarks [20]. They further captured local AU-related characteristics via spatial attention mechanism [21]. The authors of [22] proposed Geodesic Guided Convolution (GeoConv) for AU recognition by embedding 3D manifold information into 2D convolutions in which the convolutional kernel is weighted by geodesic distances on the 3D facial surface. In an attempt to assess the effectiveness of 2D and 3D CNNs in human action recognition task, the authors of [23] evaluated these networks in hand gesture recognition task.

In a separate line of research, the encoder-decoder models have been employed in this context. [24] used graph convolutional networks (GCN) for AU relation modeling. They used auto-encoders to extract latent representation of AU-related regions to be fed to GCN for modeling AU relationships. In [25], a deep structured inference network (DSIN) for AU recognition is proposed. This structure passes information obtained from extracted image features and the structure inference between predictions straightforwardly to capture the relationship between AUs. The authors of [26] proposed the AU semantic relationship embedded representation learning (SRERL) framework that first extracts global feature maps over the whole face image. Then, they process cropped features from the global feature maps, separately. Finally, they used gated graph neural networks (GGNN) to capture correlations among AUs. A Meta Auxiliary Learning method (MAL) is proposed in [27] in which adaptive weights are used for learning facial expression.”.

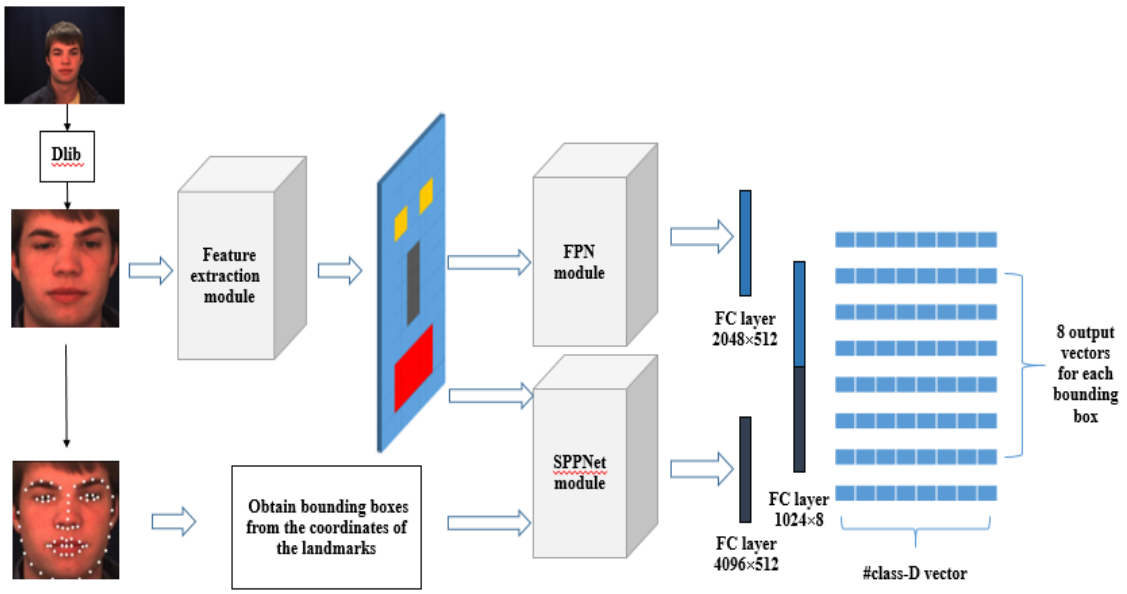

Figure 1. Outline of the proposed framework. The extracted feature maps from the initial and middle layers of ResNet-101 are fed to the FPN and SPPNet modules. Coordinates of the bounding boxes are also used by the SPPNet module. FPN and SPPNet modules work together as core modules for AU recognition.

3. Proposed Method

3.1. Overview

Most of the existing methods use only local or global features. To overcome this problem, we propose a novel architecture that covers the shortcomings of other methods. The main idea behind our proposed architecture is using both the local and global information extracted from the input data. In other words, we identify action units from locally segmented face regions while analyzing the whole face simultaneously. We combine the outcome of the two processing flows to recognize facial expressions. To this end, we use SPPNet [28] and FPN [29] **in our architecture**. Our proposed framework, as shown in Figure 1, consists of three parts: the initial convolutional layer, the SPPNet module, and the FPN module.

In the convolutional layer, we first crop faces from the images using 68 landmark points. Then, following the [18] approach we use "expert prior knowledge" to extract the coordinates of the face regions of interest (RoIs). As shown in Figure 2, this yield to eight bounding boxes each containing specific regions of face where the desired AUs happen. After that, we employ a ResNet-101 [30] and extract the feature map from each face image. In order to avoid overfitting of the model, we freeze conv1-res4 layer.

Considering the fact that the sizes of bounding boxes are not fixed for different faces, we use RoI pooling layer to fix the sizes of the feature maps obtained from conv1-res4. For this step, we map the coordinates of the bounding boxes from the original image to the feature maps and extract them. This feature maps that belong to the bounding boxes are fed to the SPPNet, and the original feature maps that belong to the whole image face are fed to the SPPNet and the FPN modules. The outputs of the two modules are concatenated, and the final fully-connected (FC) layer's output is treated as each class probability. Finally, because the prediction was at the RoI level and belonged to each bounding box, we returned the prediction results to the image level by merging the hit of each AU or AUs in each box. The details of SPPNet and the FPN modules are explained in the sequel.

3.2. SPPNet module

One of the important properties of SPPNet is the use of multi-level spatial bins. It has been shown in [28] that this architecture is robust to object deformations. In our problem, each extracted face RoI is different in scale and level of deformation, e.g., eye regions are small and have subtle deformation compared to the other regions. Therefore, in the SPPNet module, we first import the feature map into two different branches. In the lower branch, the feature maps are given to Res5, and in the other one we configure the 4-level pyramid pooling {7×7, 3×3, 2×2, (1×1) ×2} with the total of 64 bins. In order to reduce the number of channels a 1×1 Conv is used.

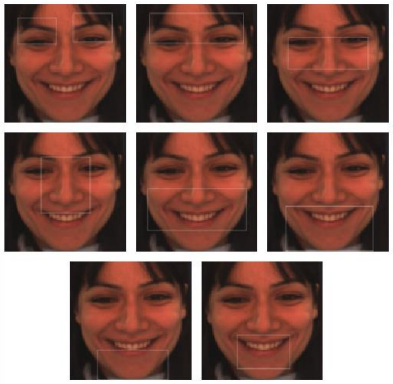

Figure 2. Eight bounding boxes for each part of the face are defined. One or more AUs may occur in each bounding box.

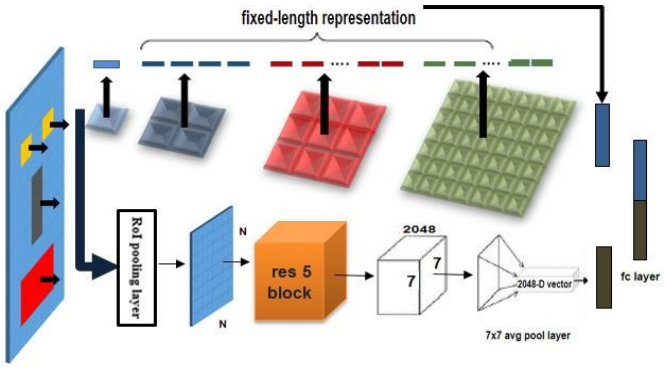

Figure 3. The SPPNet module, in which feature maps that are extracted from the initial and middle layers of ResNet101 (conv1-res4) are imported into two branches and then concatenated together and fed to a FC layer.

Finally, the outcome of the two branches are concatenated and fed to a fully connected layer. Details of the SPPNet module are illustrated in Figure 3.

3.3. FPN Module

In this module, first the low-level and high-resolution features (obtained from the previous layers) are transformed to high-level and low-resolution features using a Res5 block**.** Input and output of the Res5 block are feature maps with $1024\times M\times N$, and $2048\times M/2\times N/2$ respectively. Then the two set of features (before and after Res5) are combined to get more convenient representation. To this end, we first up-sample the output of Res5 by a factor of 2, because of the output of Res5 reduced by a factor of 2. Since the two feature sets have different dimensionalities, we fixed their dimensionality using convolutional layers, and pass both sets of features through 1×1 convolution layers. Then, the feature sets are added, and a 3×3 convolution layer is applied to prevent the aliasing effect of upsampling [29].

Since the face images are divided into eight separate regions in SPP module (see Figure 2), we replicate each feature map eight times, i.e., for each bounding box we consider a feature map of the whole face.

Elements of these eight feature maps are separately element-wise multiplied by a set of learnable coefficients as follows

$$F(x_{ij}) = \sum_{k=1}^{N} \sum_{l=1}^{C} w_{ij} * x_{ij} \qquad (1)$$

where all the w_{ij} coefficients are set to initial value of 0.01, N is set to 8, C is the number of channels, and i,j are the coordinates of the feature map elements. The details of the FPN module are illustrated in Figure 4. This way, more attention is paid to the parts of the whole face feature maps for each bounding box that convey more informative features. At the final stage of the FPN module, these feature maps are fed into the FC layer.

4. Experiments

In this section, the dataset and the experimental setup are presented first. Details of the evaluations and comparative results are provided afterwards.

4.1. Dataset

Our proposed model is evaluated on the publicly available dataset DISFA [31]. This dataset contains 54 videos, where 27 of them were captured from the left and the rest were recorded from the right side of the subject's faces. Twenty-seven young adults with diverse ethnicities participated. Each video consists of 4,485 frames, summing up to a total of about 260,000 frames. The frames are manually labeled with AU intensity on a six-point ordinal scale. Using [26, 20, 22] methods, we only considered those frames with intensities equal to or greater than 2 as positive. There are 12 AUs included in the DISFA dataset. For evaluating our framework we used 8 of them, i.e., action units 1, 2, 4, 6, 9, 12, 25, and 26.

Table 1. Results of different methods in action unit recognition task on DISFA dataset. Reported numbers are F1-frames, and bracketed and bold numbers represent the best and the second-best results, respectively.

AU	F1-frame									
	LSVM	APL	DRML	ROI-Nets	DSIN	AU R-CNN	SRERL	MAL	ARL	Our method
1	10.8	11.4	17.3	41.5	42.4	32.1	**45.7**	43.8	43.9	[47.6]
2	10	12	17.7	26.4	39.0	25.9	[47.8]	39.3	**42.1**	39.4
4	21.8	30.1	37.4	66.4	68.4	59.8	59.6	**68.9**	63.6	[70.3]
6	15.7	12.4	29	50.7	28.6	[55.3]	47.1	47.4	41.8	**52.4**
9	11.5	10.1	10.7	8.5	**46.8**	39.8	45.6	[48.6]	40.0	45.4
12	70.4	65.9	37.7	89.3	[90.4]	67.7	73.5	72.7	**76.2**	74.5
25	12	21.4	38.5	88.9	70.8	77.4	84.3	**90.6**	[95.2]	89.0
26	22.1	26.9	20.1	15.6	42.2	52.6	43.6	52.6	[66.8]	**54.8**
Avg	21.8	23.8	26.80	48.5	53.6	51.3	55.9	58.0	**58.7**	[59.2]

4.1. Implementation details

All our experiments were conducted on a computer with a GTX 1080 Ti GPU and 16 GB RAM. We used Chainer1 as our learning framework. In our processing flow, we first used Dlib2 library to get 68 landmarks for each face and then cropped the faces and resized them to 512×512 pixels. Next, we subtract the mean pixel value from all the dataset images. We augmented the dataset in a random order by horizontally flipping the input images. The size of mini-batches was set to 8. Since the backbone of our model is the same as that of AU R-CNN [18], we employed the concept of **transfer** learning, used the pre-trained model on the BP4D [32], and fine-tuned the last layers. We used Stochastic Gradient Descent (SGD) optimization algorithm and set the learning rate to 10^{-4}. The learning rate was reduced every ten epochs by a factor of 20%.

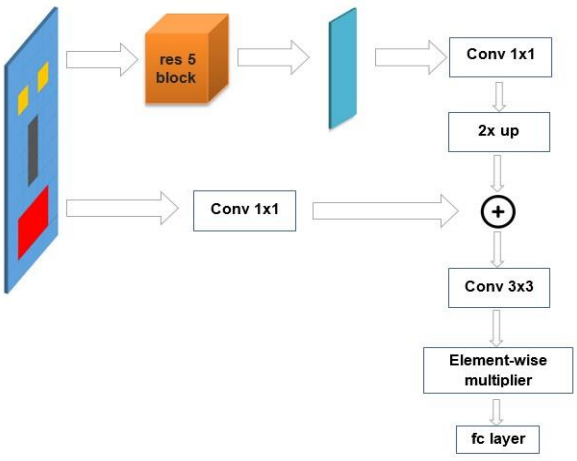

Figure 4. Details of the FPN module. Res5 block is applied to the feature maps in the top flow. These feature maps are then aggregated with the original feature maps.

[1] https://chainer.org/

[2] http://dlib.net/

Standard L^2 norm was used to regularize the network's parameters.

4.1. Evaluation Metrics

We extracted RoIs from the face images and trained our model to treat each of them as a separate bounding box (see Figure 2). In order to evaluate our method, we used the widely used accuracy measure. Because some AUs have low occurrence rates, using the accuracy measure is not enough. Therefore, we also used the F1-frame (F1-score) [33], which is commonly used in the literature and is defined using precision and recall as follows:

$$F1 = 2precision * recall/(precision + recall) \quad (2)$$

All experiments were conducted in a subject-exclusive 3-fold cross-validation scheme and accuracy and F1-frame score for all the AUs were calculated. The reported values are the average results (denoted as Avg.) over all experiments.

4.4. Results

In the following, the results of the proposed method are compared with those of similar methods on the DISFA dataset. The results presented in Table 1 and Table 2 are obtained under 3-fold cross-validation setting. The reported values in the tables are F1-frame and accuracy, respectively. Traditional methods like linear support vector machine (LSVM) [34], active patch learning (APL) [35], deep region and multi-label learning (DRML) [36], ROI adaption net (ROI-Nets) [37], DSIN [25], AU R-CNN [18], and the recent successful methods like SRERL [26] and ARL [21] were compared with our method. It should be noted that our comparison includes only those methods that use static two-dimensional image representation.

Table 2. Results of different methods in action unit recognition task on DISFA dataset. Reported numbers are accuracies, and bracketed and bold numbers represent the best and the second-best results, respectively.

AU	LSVM	APL	DRML	SRERL	ARL	Our
1	21.6	32.7	53.3	76.2	**92.1**	[94.7]
2	15.8	27.8	53.2	80.9	**92.7**	[93.5]
4	17.2	37.9	60.0	79.1	**88.5**	[88.8]
6	8.7	13.6	54.9	80.4	**91.6**	[91.9]
9	15.0	64.4	51.5	76.5	[95.9]	**95.7**
12	93.8	94.2	54.6	87.9	[93.9]	**92.1**
25	3.4	50.4	45.6	90.9	[97.3]	**93.1**
26	20.1	47.1	45.3	73.4	[94.3]	**91.0**
Avg	27.5	46.0	52.3	80.7	[93.3]	**92.6**

Table 1 shows the performance of different methods in terms of F1-score. Our proposed method outperforms other methods on average of F1-scores. Moreover, our method achieves the best classification results for AUs 1 and 4, and the second-best classification results for AUs 6, 25, and 26. It is observed from the reported results in Table 1 that the other methods also perform differently for different action units. For example, the results of the ARL [21] method are the best for AUs 22 and 26 and the second-best for AUs 2 and 12. The outcome of the same method for AUs 4, 6, and 9 is significantly low compared with that of the other methods. Our proposed method is consistently performing good for all AUs, and although not all of our results are the best, the performance of the method is comparable with that of the others. Generalization capability of our method can be the result of simultaneous utilization of local and global face features through the FPN and SPPNet modules.

The proposed method outperforms all other methods for AUs 1, 2, 4, and 6. and is the second best for others in terms of classification accuracy. As shown in Table 2, there is a significant gap between the two best algorithms (i.e. ARL and our proposed method) and the other methods. This shows the strength of the proposed architecture. On the other hand, our method and the AU R-CNN method both use ResNet-101 as the backbone of the models. However, the number of learnable parameters of our method is much less than that of AU R-CNN. This is because we freeze the initial and middle layers and don't train them. Therefore, the proposed method is more efficient than AU R-CNN. Moreover, another advantage of our method is its simplicity compared to the other methods specifically ARL.

4. Conclusion

A common approach for detecting facial expressions is to recognize different facial action units and then use their combination to identify the facial expressions. To this end, some methods use the whole face as a single object to detect and classify action units, while the others detect each action unit separately. Despite the achievements, the latter approaches are prone to misclassification because they miss some useful information. For example, features from the upper face, such as those related to eye and eyebrow gestures, can enhance the detection performance of the AUs in the lower face (e.g., AU 25, AU 26). By using the whole face images, it is possible to capture each AU's global and occurrence-related features.

The strengths of each of these two approaches inspired us to use the combination of them to complement each other and eliminate possible flaws. It is known that the constituent area (and scales) and the appearance of each AU may vary. Therefore, we used SPPNet [28] to generalize the proposed model to learn these variations. With an end-to-end trainable framework (SPP-FPNNet), we proposed to combine local features using the SPP module with global features using the FPN module to achieve an efficient approach because of less time has been spent to train the network. This is achieved by freezing the weights of the initial and middle layers of the network, and fine-tuning the last layers. Our proposed model outperforms the state-of-the-art methods for the well-known challenging DISFA dataset. By training the whole networks parameters, we expect to achieve better results. However, this imposes high computational cost.

As the future line of research, we would like to use temporal features as an integral element in detecting AUs. Moreover, using sequence modeling techniques, more specifically attention mechanism, can be a logical extension of the current work to be able to tackle the temporal dynamics in videos.

References

[1] P. Ekman and W. V. Friesen, Unmasking the face: A guide to recognizing emotions from facial clues. vol. 10, Ishk, 2003.

[2] S. Wang, H. Ding, and G. Peng, "Dual learning for facial action unit detection under non-full annotation." *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2225-2237, April 2022.

[3] C. Fabian 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 the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[4] R. Ekman, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA,1997.

[5] K. Zhao, W.-S. Chu, and A. M. Martinez, "Learning facial action units from web images with scalable weakly supervised clustering," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 2090–2099.

[6] R. Zhi, M. Liu, and D. Zhang, "Facial representation for automatic facial action unit analysis system," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 2019, pp. 1368–1372.

[7] J. J. Bazzo and M. V. Lamar, "Recognizing facial actions using gabor wavelets with neutral face average difference," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. Proceedings. IEEE, 2004, pp. 505–510.

[8] 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). IEEE*, 2006, pp. 149–149.

[9] J. Whitehill and C. W. Omlin, "Haar features for facs au recognition," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 5–pp.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1999, pp. 1150–1157.

[11] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Face and Gesture 2011. IEEE*, 2011, pp. 314–321.

[12] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131–146, 2000.

[13] 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 emotionspecified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE*, 2010, pp. 94–101.

[14] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 28–43, 2011.

[15] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[16] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, no. 83, pp. 2491–2521, 2008.

[17] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Joint facial action unit detection and feature fusion: A multi-conditional learning approach," *IEEE transactions on image processing*, vol. 25, no. 12, pp. 5727–5742, 2016.

[18] C. Ma, L. Chen, and J. Yong, "AU R-CNN: Encoding expert prior knowledge into RCNN for action unit detection," *Neurocomputing*, vol. 355, pp. 35–47, 2019.

[19] M. Savadi Hosseini and F. Ghaderi, "A hybrid deep learning architecture using 3D CNNs and GRUs for human action recognition," *International Journal of Engineering*, vol. 33, no. 5, pp. 959–965, 2020.

[20] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action
unit detection and face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 705–720.

[21] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *IEEE Transactions on Affective Computing*, vol. 13, 2019, pp. 1274–1289.

[22] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, and J. Zheng, "Geoconv: Geodesic guided convolution for facial action unit recognition," *Pattern Recognition* 122 (2022): 108355.

[23] M. Kurmanji and F. Ghaderi, "Hand gesture recognition from RGB-D data using 2D and 3D convolutional neural networks: a comparative study," *Journal of AI and Data Mining*, vol. 8, no. 2, pp. 177–188, 2020.

[24] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *International Conference on Multimedia Modeling. Springer*, 2020, pp. 489–501.

[25] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11216, 2018, pp. 298–313.

[26] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8594–8601.

[27] Y. Li, S. Shan, "Meta auxiliary learning for facial action unit detection" *IEEE Transactions on Affective Computing*, vol. 19, 2021, pp. 14–17.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[29] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[32] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[33] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction. IEEE*, 2013, pp. 245–251.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

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

[36] 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*, 2016, pp. 3391–3399.

[37] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multilabeling learning and optimal temporal fusing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1841–1850.

شناسایی واحد حرکتی چهره بهبودیافته با استفاده از ویژگی های محلی و سراسری چهره

امین رحمتی سردشت و فؤاد قادری*

آزمایشگاه تعامل انسان و کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران.

ارسال ۲۰۲۲/۰۷/۱۷؛ بازنگری ۲۰۲۲/۱۰/۱۶؛ پذیرش ۲۰۲۳/۰۱/۱۰

چکیده:

هر هیجان چهره شامل یک یا چند واحد عملی است که روی چهره ظاهر می‌شود. بنابراین، شناسایی واحد عملی معمولا برای افزایش راندمان تشخیص هیجان چهره استفاده می‌شود. شناسایی تغییرات ظریف در چهره زمانی که واحدهای حرکتی خاصی رخ می‌دهند، مهم است. در این مقاله، ما یک معماری پیشنهاد می‌کنیم که از ویژگی‌های محلی استخراج‌شده از نواحی خاصی از چهره و در عین حال استفاده از ویژگی‌های سراسری گرفته‌شده از کل چهره استفاده می‌کند. برای این منظور، ماژول‌های SPPNet و FPN را با هم ترکیب می‌کنیم تا یک شبکه سرتاسر برای تشخیص واحد عملی چهره طراحی کنیم. ابتدا نواحی مختلف از پیش تعریف شده چهره شناسایی می‌شوند و در مرحله بعد، ماژول SPPNet تغییر شکل‌ها را در نواحی شناسایی شده ثبت می‌کند. ماژول SPPNet بر روی هر ناحیه چهره به طور جداگانه تمرکز می‌کند و نمی‌تواند تغییرات احتمالی در سایر نواحی صورت را در نظر بگیرد. به موازات آن، ماژول FPN ویژگی‌های سراسری مربوط به هر یک از نواحی چهره را پیدا می‌کند. با ترکیب این دو ماژول، معماری پیشنهادی می‌تواند هم ویژگی‌های محلی و هم سراسری چهره را ثبت کند و راندمان شناسایی واحد عملی را افزایش دهد. نتایج تجربی روی مجموعه‌داده DISFA اثربخشی روش ما را نشان می‌دهد.

کلمات کلیدی: شبکه‌های عصبی پیچشی، لایه ادغام هرم فضایی، شناسایی هیجانات چهره، واحد حرکتی.