Multi-branch Multi-Scale Attention Network for Facial Expression Recognition (FER) in-the-Wild

Chakrapani Ghadai and Dipti Patra

National Institute of Technology, Rourkela, India
520ee7014@nitrkl.ac.in
dpatra@nitrkl.ac.in

Abstract. The challenges in facial expression recognition (FER) is mostly caused by high intra-class variations, subtle inter-class visual changes, and smaller datasets. The intra-class and inter-class variations suffer big from the pose, illumination, or partial occlusion in the real world, which degrade the performance of FER significantly. Multi-scale and attention-based networks are widely used to address these challenges. In most of the previous approaches, lower-level features at smaller scale progress towards higher levels to construct features at larger scales, or convolutions at different resolutions are used for multi-scale feature representations. The used methods have increased depth, but lacked width and are inadequate in representing features at granular levels to precisely capture important facial expression features. Here, we introduce a novel multi-branch multi-scale attention network (MSA-Net) for FER. MSA-net is a deeper and wider network and it extracts multi-scale features at different receptive fields in a parallel network structure. Moreover, to improve the effective receptive field and extract diverse features, different kernel sizes in each parallel branch are used. Further, to focus on important regions, and make the features more discriminating multi-scale features are passed through attention networks. MSA-Net can extract sufficiently diverse attention-enhanced multi-scale features from different parallel paths, this can lessen the effect of intra-class and inter-class variations due to external factors. Further, features at different receptive fields from each parallel path are combined together to reduce the effect of pose and partial occlusion. The experimental findings reveal that the suggested method achieves competitive results on widely used in-the-wild public datasets.

Keywords: Facial expression recognition · CNN · Muti-scale · Attention · receptive field · kernel size.

1 Introduction

The facial expression conveys one's emotional state of mind, and it is one of the most powerful nonverbal communication methods for social interactions. The research in FER is growing due to its broad applications in many domains e.g. human-computer interaction [25], modern driver aid systems [28], and medical diagnosis [1] etc. FER system assigns a category to a still image or a video out of many basic emotions. FER from still images is called static FER and FER from a video is called dynamic FER. Further, FER can be divided into controlled and in-the-wild depending upon the dataset used. Recently, FER on laboratory-controlled datasets has achieved significant results due to all frontal images taken under controlled laboratory conditions. However, FER on in-the-wild datasets is challenging due to large intra-class variations, very small inter-class visual differences, and complexity in the representation of subtle local variations caused by muscle movements of dynamic face regions.

Over the years, researchers followed different approaches to address the above challenges in FER. Traditional methods used gabor filter[21], histograms of oriented gradients (HOG) [4], scale-invariant feature transform (SIFT) [23] etc. for FER. These methods use shallow learning for feature extraction, are subject to underfit and the performance of these methods are not encouraging. Recently, much work used deep learning especially convolutional neural network (CNN) [19, 22] in FER which significantly improved performance over traditional shallow learning methods. The methods used deeper CNNs, which have better generalization capability, but can be easily overfitted in presence of external factors e.g. pose, illumination, occlusion, etc. To focus on emotion related regions many methods used attention networks in CNN backbone [15] for FER in-the-wild. Further, some methods used DenseNet, a variant of inception [11], and pyramid [24] as multi-scale feature extractors, and multi-scale features serve as context to attention networks for learning more discriminating features. However, many of the existing methods either use convolutions at different resolutions in parallel paths or use a single path for multi-scale feature extraction. In single path multi-scale networks, lower-level features at smaller receptive fields progress towards higher levels to construct features at a larger receptive field. Both the single path based and multi-resolution-based methods lack diversity and a larger effective receptive field in feature learning. To overcome some of the above challenges a novel multi-branch multi-scale attention network (MSA-Net) for FER in-the-wild is proposed.

The main contributions are summarized as follows:

1. Multi-branch multi-scale attention network is constructed to learn features at different receptive fields and the feature maps of each branch are fused together, which can reduce the effect of pose and partial occlusion.

2. The multi-scale block used extracts multi-scale features by combining features at a different receptive field in a hierarchical fashion within a single multi-scale block, which can lessen the FER system's vulnerability to inter-class and intra-class variations due to external factors.

3. The attention module is used in the deep layer of each branch of MSA-Net to focus on important regions and make the expression-related features more discriminating.

The remaining part of this paper is structured as follows. Section 2 provides an overview of the proposed MSA-Net. Section 3 discusses the experimental details, results and discussion. In Section 4, we finally give concluding remarks.

2 Proposed Method

2.1 A Framework Overview

Fig. 1 shows the overall framework of the proposed MSA-Net. MSA-Net consists of feature pre-extractor, parallel multi-scale, and attention network blocks. The feature pre-extractor extracts middle-level feature maps of size 128x28x28 through the use of one 2D convolution layer, a max pooling layer, and two layers of ResNet-18 [12]. Then, middle-level feature maps are passed through three parallel branch networks; each of the parallel branches consists of a multi-scale module and an attention module. Inside the multi-scale module, two multi-scale blocks are connected in series. 7x7, 5x5, and 3x3 convolutions are used in multi-scale blocks of the first, second, and third branches respectively for finding multi-scale features at different receptive fields. Finally, the feature maps from three branches are fused at the feature level to obtain the FER results.

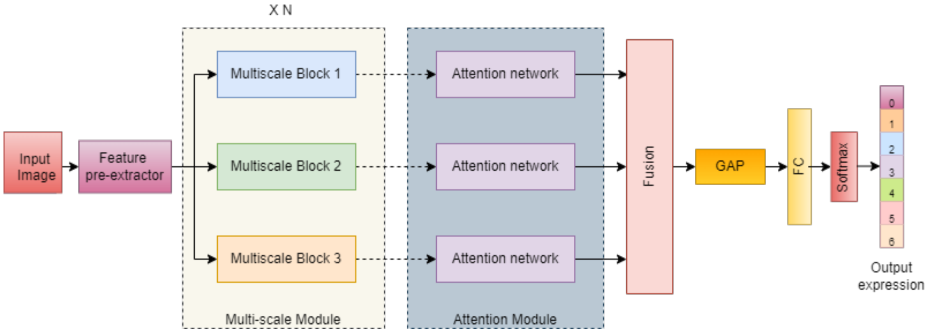


Fig. 1: The block diagram of MSA-net showing different components.

2.2 Multi-scale Module

Two multi-scale blocks connected in series in each parallel path constitute a multi-scale module. The designed multi-scale block is inspired by Res2Net [9]. Each multi-scale block uses bottleneck blocks in a hierarchical-like manner within a single residual block to learn features at different receptive fields. Moreover, convolution with higher kernel sizes is biased towards shape, and convolution with smaller kernel sizes is biased towards texture [6]. So, multi-scale blocks in each branch use different kernel sizes to make the multi-scale feature maps diverse. Fig. 2 shows the basic multi-scale block, 1x1 convolution layers are used for down sampling and up sampling, and the no of input and output channels remain the same in the bottleneck convolution layer. This arrangement reduces the computation overhead due to the larger kernel size. Let X be the feature maps obtained after k × k convolution. The feature map X is split into n equal feature map subsets along the channel axis. Let X_i be the feature map subsets, where $i \in \{1, 2, ..., n\}$. Hence, the feature map subsets $X_i \ \forall \ i$ have the same spatial size, and the number of channels of is $\frac{1}{n}$th of the total number of channels of X. Moreover, the feature map subsets $X_i \ \forall \ i$ are processed in three parallel multi-scale blocks. Let the m × m convolution process applied on each of X_i is denoted by $P_i^p(.)$,where $p \in \{1, 2, 3\}$, p denotes the position of the multi-scale block. Let Y_i^p denotes the output of $P_i^p(.)$. Therefore, the output Y_i^p for each feature map subset X_i can be written as:

$$Y_i^p = \begin{cases} P_i^p(X_i), & i = 1 \\ & & ;1 < p \leq 3 \\ P_i^p(X_i) + Y_(i-1)^p, & 1 < i \leq n \end{cases} \tag{1}$$

Let's consider the first multi-scale block i.e. $p = 1$, notice that each 7x7 convolutional operator $P_i^p(.)$ receive feature information from all feature subsets $X_j : j < i$. Each time a feature subset X_i passes through a 7x7 convolutional operator, the output feature map attains a larger receptive field than X_j. Moreover, the second and third multi-scale blocks have 5x5 and 3x3 convolutional operators respectively. Hence, the first multi-scale block has a higher effective receptive field than the second multi-scale block, and the second multi- scale block has a higher effective receptive field than the third multi-scale block.

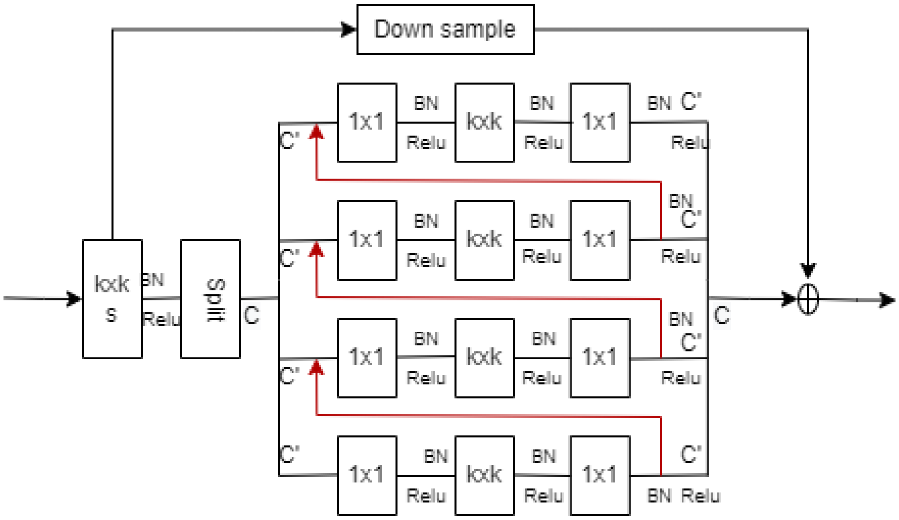


Fig. 2: Multi-scale block, k=7, 5, 3 for multi-scale block 1, 2, 3 respectively.

2.3 Attention Module

Let the feature maps of each parallel path obtained after the multi-scale module be denoted as $F^p \in \mathbb{R}^{H \times W \times C}$, where $p \in \{1, 2, 3\}$. Then, each of the feature maps are passed through an attention network. A convolutional block attention module (CBAM) [29] is used as the attention network. In CBAM, attention feature maps are extracted sequentially along the channel and spatial dimensions, then the refined adaptive feature maps are obtained by multiplying attention maps with the input feature maps. For each parallel path, the attention network computes channel attention map $M_c^p \in \mathbb{R}^{1 \times 1 \times C}$ and spatial attention map $M_s^p \in \mathbb{R}^{H \times W \times C}$. Therefore, the final refined output of attention networks from each path can be formulated as: $F_r^p = M_s^p(F) \circledast (M_c^p(F) \circledast F)$ here $\circledast$ denotes element-wise multiplication. The channel and spatial attention maps are scaled along spatial and channel directions respectively for conducting multiplication. Finally, the attention feature maps from the parallel branches are fused and let it be denoted as F_r.

2.4 Fusion Strategy and loss function

The feature maps from the parallel branches are fused using simple concatenation operations along the channel axis. The global average pooling (GAP) operation is performed on the final fused attention feature maps to obtain a vector,v. The end-to-end model is trained by the cross-entropy loss function as given below.

$$L = -\frac{\beta}{N} \sum_{0}^{N-1} \log \frac{e^{w_i^T v_i + b_i}}{\sum_0^{C-1} e^{w_j^T v_j + b_j}} \tag{2}$$

Where N is the input batch size; C is the number of expression categories; W and b are the weight and bias terms of the FC layer; v_i is the input to the FC layer at i^{th} sample. β is the hyperparameter of the loss function.

3 Results and Discussion

3.1 Datasets and Implementation details

Two popular FER in-the-wild datasets, RAF-DB [16] and Affectnet [20], are used for conducting the experiments. RAF-DB contains 29672 images annotated with basic or compound, 12264 as training samples, and 3061 as testing samples from basic images taken. In the Affectnet dataset, 283,901 images as training data and 2992 images as test data are selected. We conducted an experiment for seven basic expression categories. In both datasets, the officially aligned data samples are used directly, and then they are resized to 224x224 pixels. Simple data augmentation techniques are employed to extract the random crops (central, corner, and horizontal flips). We trained the model in NVIDIA Geforce RTX 3060 GPU using an SGD optimizer with a momentum of 0.9, batch size of 32, and initial learning rate of 0.1 decayed by a factor of 10 for every 10 epochs. Hyper-parameter β is empirically set as 0.6. Further, we used four numbers of workers in multi-process data loading to speed up the training process.

3.2 Comparison of results

We present the experiment results in Table 1 and Table 2 for RAF-DB and Affectnet datasets, respectively. The proposed MSA-net method outperforms some of the state-of-the-art methods with a recognition accuracy of 86.32 and average accuracy of 79.88 on the RAF-DB dataset. Similarly, it achieves a recognition accuracy of 63.61 and average accuracy of 57.63 on the Affectnet dataset. Fig. 3 presents the confusion matrices of the RAF-DB and AffectNet.

Table 1: Comparison of results on RAF-DB dataset.

Method	Backbone	Year	Acc. (%)
RLPS [14]	10-layer DCNN	2020	72.89
Sadeghi & Raie [21]	-	2019	76.23
SCN [26]	ResNet-18	2020	78.31
pACNN [18]	Densenet	2018	83.27
ALT [8]	-	2019	84.50
gACNN [17]	VGG-16	2019	85.07
Proposed method	Manually designed	2022	86.32

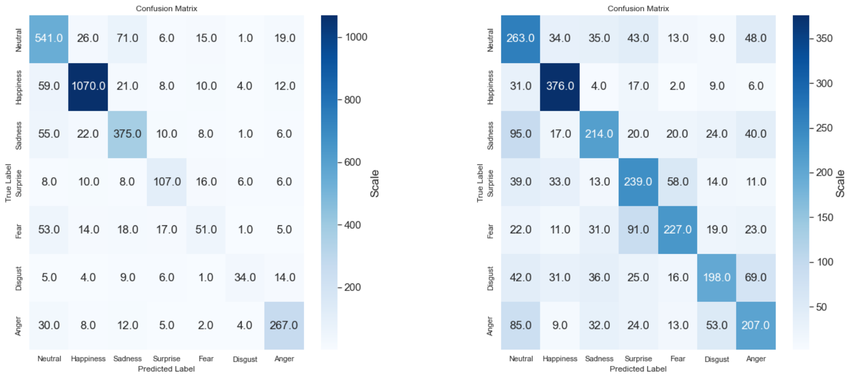
Table 2: Comparison of results on Affectnet dataset.

Method	Backbone	Year	Acc. (%)
pACNN [18]	Densenet	2018	55.33
gACNN [17]	VGG-16	2019	58.78
LDL-ALSG [2]	ResNet-50	2020	59.35
VGG-FACE [13]	VGG-16	2020	60.00
FMPN [3]	Inception-V3	2019	61.52
OADN [5]	Manually designed	2020	61.89
DDA-Loss [7]	DCNN	2020	62.34
LLHF [10]	VGG	2018	63.31
Proposed method	Manually designed	2022	63.61

3.3 Ablation Analysis

Ablation analysis on RAF-DB, Affectnet and Pose-AffectNet [27] datasets are performed to evaluate the effectiveness of each component used in the proposed method. Table 3 shows the analysis of the effectiveness of the multi-scale module and attention module in one branch network structure, Resnet-18 is considered as one branch baseline, and it's also observed that MSA-net gives the best result with three parallel branch structures.

Multi-scale Structure (MS) The multi-scale structure in one branch network of MSA-Net without attention module is studied. It is observed from Table 3 that MS improves the baseline by 2.1%, 3.08%, 2.61%, 2.43% on RAF-DB, AffectNet Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively.



(a) RAF-DB

(b) AffectNet

Fig. 3: Confusion matrices

Table 3: Evaluation of each component on RAFDB, Affectnet, and pose-affectnet

Methods	RAF-DB	Affectnet	Pose-AffectNet≥ 30°	Pose-AffectNet ≥ 45°
Baseline	82.59	59.4	53.55	52.39
MS	84.33	61.23	54.85	53.66
AS	85.36	62.11	55.77	54.81

Attention Structure (AS) The attention module is added to a multi-scale structure in one branch network. It is clear that the attention network embedded with a multi-scale module significantly improves accuracy by 1.22%, 1.46%, 1.68%, and 2.14% on RAF-DB, AffectNet Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively. Hence, the Multi-scale module extracts diverse multi-scale features and the attention network further makes features more discriminating.

4 Conclusion

A novel multi-branch multi-scale attention network (MSA-Net) for in-the wild facial expression recognition is proposed to address the issues of both intra-class and inter-class variations due to external factors like occlusion and pose. The multi-scale module learns features at different receptive fields, which can reduce the effect of pose and partial occlusions in the inference process. The attention module can focus on the important part, neglect other parts, and extract features at granular levels which can further reduce the susceptibility of the network toward subtle expression-related variations. The proposed method MSA-Net has achieved competitive accuracy with 86.32% and 63.61% on RAF-DB and AffectNet respectively. MSA-Net presents a coarse feature learning model focusing mainly on large intra-class and inter-class variations due to pose, and partial occlusions. In future work, we will investigate subtle intra-class visual differences based on feature diversification and adaptive learning to improve the performance of the model.

References

1. Canal, F.Z., Müller, T.R., Matias, J.C., Scotton, G.G., de Sa Junior, A.R., Pozzebon, E., Sobieranski, A.C.: A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences **582**, 593–617 (2022)
2. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13984–13993 (2020)
3. Chen, Y., Wang, J., Chen, S., Shi, Z., Cai, J.: Facial motion prior networks for facial expression recognition. In: 2019 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2019)
4. Dahmane, M., Meunier, J.: Emotion recognition using dynamic grid-based hog features. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 884–888. IEEE (2011)
5. Ding, H., Zhou, P., Chellappa, R.: Occlusion-adaptive deep network for robust facial expression recognition. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–9. IEEE (2020)
6. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11963–11975 (2022)
7. Farzaneh, A.H., Qi, X.: Discriminant distribution-agnostic loss for facial expression recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 406–407 (2020)
8. Florea, C., Florea, L., Badea, M.S., Vertan, C., Racoviteanu, A.: Annealed label transfer for face expression recognition. In: BMVC. p. 104 (2019)
9. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence **43**(2), 652–662 (2019)
10. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019)

11. Hardjadinata, H., Oetama, R.S., Prasetiawan, I.: Facial expression recognition using xception and densenet architecture. In: 2021 6th International Conference on New Media Studies (CONMEDIA). pp. 60–65. IEEE (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Deep neural network augmentation: Generating faces for affect analysis. International Journal of Computer Vision **128**(5), 1455–1484 (2020)
14. Li, H., Xu, H.: Deep reinforcement learning for robust emotional classification in facial expression recognition. Knowledge-Based Systems **204**, 106172 (2020)
15. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based cnn for facial expression recognition. Neurocomputing **411**, 340–350 (2020)
16. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
17. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing **28**(5), 2439–2450 (2018)
18. Li, Y., Zeng, J., Shan, S., Chen, X.: Patch-gated cnn for occlusion-aware facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2209–2214. IEEE (2018)
19. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Transactions on Instrumentation and Measurement **70**, 1–12 (2020)
20. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
21. Sadeghi, H., Raie, A.A.: Human vision inspired feature extraction for facial expression recognition. Multimedia Tools and Applications **78**(21), 30335–30353 (2019)
22. Sadeghi, H., Raie, A.A.: Histnet: Histogram-based convolutional neural network with chi-squared deep metric learning for facial expression recognition. Information Sciences **608**, 472–488 (2022)
23. Soyel, H., Demirel, H.: Facial expression recognition based on discriminative scale invariant feature transform. Electronics letters **46**(5), 343–345 (2010)
24. Vo, T.H., Lee, G.S., Yang, H.J., Kim, S.H.: Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access **8**, 131988–132001 (2020)
25. Wang, H.H., Gu, J.W.: The applications of facial expression recognition in human-computer interaction. In: 2018 IEEE international conference on advanced manufacturing (ICAM). pp. 288–291. IEEE (2018)
26. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020)
27. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)
28. Wilhelm, T.: Towards facial expression analysis in a driver assistance system. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–4. IEEE (2019)

29. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)