Facial Expression Recognition Using Residual Masking Network

Luan Pham†, The Huynh Vu †, Tuan Anh Tran *†
† Research Department-Cinnamon AI, Viet Nam
* Faculty of Computer Science & Engineering, Ho Chi Minh City-University of Technology (HCMUT),
268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
Corresponding author
phamquiluan@gmail.com, vuthe_huynh@yahoo.com, trtanh@hcmut.edu.vn

***Abstract*—Automatic facial expression recognition (FER) has gained much attention due to its applications in human-computer interaction. Among the approaches to improve FER tasks, this paper focuses on deep architecture with the attention mechanism. We propose a novel Masking Idea to boost the performance of CNN in facial expression task. It uses a segmentation network to refine feature maps, enabling the network to focus on relevant information to make correct decisions. In experiments, we combine the ubiquitous Deep Residual Network and Unet-like architecture to produce a Residual Masking Network. The proposed method holds state-of-the-art (SOTA) accuracy on the well-known FER2013 and private VEMO datasets.**

***Index Terms*—Facial Expression Recognition, Masking Idea, Residual Masking Network**

I. Introduction

Facial expression is one of the means of non-verbal communication, which accounts for a significant proportion of human interactions [1], [2]. It can be represented as discrete states (such as anger, disgust, fear, happiness) [3] based on cross-culture studies [4]. Human emotions are sometimes mixed together in specific time and space conditions. However, ignoring the intricate emotions intentionally created by humans, the primary emotion is still widespread due to its intuitive definition. Like most other FER methods, our approach focuses on recognizing six facial emotional expressions (proposed by Ekman et al. [3]) and the neutral state on the static image and ignoring the temporal relationships [4].

The FER task presents several challenges, especially in-the-wild settings due to the difference between inter-subject and intra-subject. For inter-subject variations, faces of individuals vary depending on different gender, ages, or ethnic groups. On the other hand, intra-subject changes include occlusions, illumination, and variations of head poses. Despite challenges, research in FER has drawn much attention, leading to several practical applications in human-computer interaction systems and data analytics [4], [5] [6].

In recent times, with the popularity of deep learning, especially convolutional neural networks, deep features can be automatically extracted and learned for a good facial expression recognition system [4]. However, in the case of facial expressions, it is noteworthy to mention that much of the cues come from a few facial regions like eyes, mouth. In contrast, other areas have little contribution to the output, e.g. hair, jawline. Methods are trying to focus on these essential regions by using an intermediate feature named facial landmark [7]. Facial landmark detection can get striking results in the lab-controlled condition. Still, in a noisy environment, they usually do not perform very well due to the variants in head poses, illumination, etc., (see Figure 1). Attention mechanism for the image classification problem in recent studies has been developed to increase the performance of the convolution neural network by focusing on tiny details [8]–[10]. Besides, in image segmentation problems, top-down bottom-up like architectures can keep useful information in pixel-level effectively [11], [12]. From these points, we propose a novel Masking Idea. This idea uses a Unet-based localization network to refine input feature maps and generates output feature maps that contain attention to some areas of the input feature maps. Each Masking Block is a small variance of the U-net network, which enables the Residual Masking Network to focus on crucial spatial

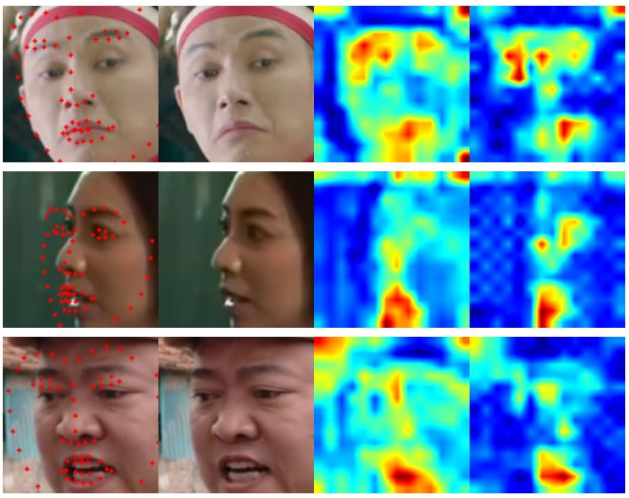


Fig. 1: Example of landmark detection and features of Masking Block as follows: landmark detection, original image, feature map before the 3^{rd} Masking Block, feature map after the 3^{rd} Masking Block.

information and to make correct facial expression classification.

In this research, we provide three significant contributions: Proposing a novel Masking Idea to boost the performance; build a Residual Masking Network for facial expression recognition; create a new dataset named VEMO. Our works are available online at GitHub [1].

The remaining parts of this paper are organized as follows: Section II provides a brief review of previous related studies, Section III describes the methods used for the network architecture, Section IV includes a discussion about experimental results, and the paper is concluded in Section V.

II. Related works

Even though the deep learning approaches recently show the efficiency in automatic facial expression recognition, traditional machine learning methods remain prevalent. These methods are still active in many cases and continue to be developed. Traditional FER approaches have been using handcrafted features and experiments mostly on lab-controlled datasets. Features taken in places such as the eyes, nose, and mouth must be consistent across the image. Some well-known feature extractions can be listed as Local Binary Patterns (LBP) [13], LBP on Three Orthogonal Planes (LBP-TOP) [14], Non-negative Matrix Factorization (NMF) [15], Sparse Learning [16].

In most of the traditional approaches, the first step in practice is to detect the position of the face, then, extract geometric features [16], appearance features [17], or both [18] to generate specific vectors to the model. These methods are generally quite complex, requiring a lot of technical manipulation. The characteristic analysis becomes extremely challenging as the data is massive. These methods often face the problem in natural or noise environments where landmark detection is difficult.

Leading to higher accuracy compared to traditional handcrafted features, recent static image FER has followed deep learning approaches. Lots of training images are required for deep learning-based networks to prevent overfitting. The introduction of more large scale datasets such as EmotioNet [19], AffectNet [20], ExpW [21], FER2013 [22], and the increasing computational power (such as GPUs, TPU) enabled more applications of these approaches. Many CNN architectures were applied for FER to increase the expressiveness of feature representation. Yao et al. [23] introduced the HoloNet by incorporating concatenated RELU (CRELU) [24] and inception-residual blocks [25] to the existing Resnet structure [26] to increase the network's depth and improve the multi-scale learning. In another design proposed by [27], three supervised blocks were applied to raise the degree of supervision of the Resnet network [26].

The Network ensemble is another strategy being applied for FER to improve its accuracy. Individual networks are ensembled by concatenating their features [28] or taking an average of their output predictions [29]. For the ensembling to take effect, the network should have sufficient diversity by being trained on different training data, having varied architectures, parameters, size of filters, or several network layers. Hamester et al. [30] combined convolutional neural networks (CNNs) with a convolutional auto-encoder (CAE) for architecture ensemble.

In traditional CNNs for the classification tasks, the loss is often applied to keep features of a different class apart. To increase the discrimination among other facial emotional expressions, Cai et al. [31] proposed the island loss to decrease the variations among intra-class while expanding the difference among inter-class simultaneously.

Combining traditional methods with deep learning can be an effective solution to this problem. A number of combined methods are proposed by [32] [16] [33]. As we can see in Table IV, their results are quite well.

The main problem, however, is still the complexity of technical manipulation. Besides, those methods are usually designed optimally for a specific data set (or a particular target), which leads to low re-usability. Summarizing state of the art in FER, We can draw the following key points as a basis for developing our network. Firstly, the accuracy of detecting essential facial areas contributes mainly to the improvement of the classification accuracy. Secondly, Facial expressions are determined based on the combination of some facial regions such as the eyes, nose, mouth [7]. Several traditional methods extracted these facial areas based on facial landmarks. However, the landmark detection worked well mostly on lab-controlled datasets but not in-the-wild datasets due to intra-subject variations such as occlusions, illumination, and variations of head poses (see Figure 1). Thirdly, For CNN-based FER approaches, the localization of facial areas can be observed through intermediate layers of CNN (the third and fourth column of Figure 1). Also, attention mechanisms were often applied to CNN-based approaches to improve a network's concentration on relevant information and ignore unnecessary ones.

Wang et al. [10] developed the Attention Module including a trunk branch and a mask branch with the argument that the trunk branch perform the feature processing and the mask branch will produce the same size mask that softly weights the output features of trunk branch. Analysis of this idea, we have two proposals: It will be better if the output of the mask branch can score the importance of the output activation maps of the trunk branch. And, a more in-depth network might achieve a better result in localizing the vital score of the feature maps, [12].

The masking idea argues that a localization network might help to refine the tensors by producing its importance weights, which supports the learning process to focus on what it deems necessary. The loss function would track the refinement. This intuition is the same as the way the U-net network receives a biological image and returns a segmentation mask.

[1] http://github.com/phamquiluan/ResidualMaskingNetwork

III. Proposed method

A. Overview

The main flow of the proposed method is the Residual Masking Network illustrated in Figure 2. This network contains four main Residual Masking Blocks (Resmasking Blocks). Each Residual Masking Block, which operates on different feature sizes, contains a Residual Layer and a Masking Block (see Table I).

An input image of size 224×224 will go through the first 3×3 convolutional layer with stride 2 before passing a 2×2 max-pooling layer, reducing its spatial size to 56×56. Next, the feature maps obtained after the previous pooling layer are transformed by the following four Residual Masking Blocks with generated features maps of four spatial sizes, including 56×56, 28×28, 14×14, and 7×7. The network ends with an average pooling layer and a 7-way fully-connected layer with softmax to produce outputs corresponding to seven facial expression states (6 emotions and one neutral state).

B. Residual Masking Block

In this research, we propose the Masking Block, which performs the scoring operation. Then, input feature maps of the Masking Block and its outputs are directly combined. We remove the trunk branch and re-use Resnet34 [26] as the backbone.

We design Residual Masking Block containing a Residual Layer and a Masking Block, with the former being in charge of feature processing and the latter producing the weights for the corresponding feature maps, as in Figure 2.

Given an input feature map $F \in R^{C \times W \times H}$, firstly, F will go through Residual Layer R (Figure 2a) to produce the coarse feature map $F_R = R(F)$, $F_R \in R^{C' \times W' \times H'}$. Then, the same size activation map F_M having a value in range $[0, 1]$ are calculated via Masking Block by the formula $F_M = M(F_R)$. Finally, the refined feature map - output of Residual Masking Block will be yielded by the Formula 1. This way, we assume that F_M would be more convenient to score the element-wisely importance of the input feature map F_R than the changed one [10].

$$F_N = F_R + F_R \otimes F_M, \tag{1}$$

Where F_R is a transformed feature map of F via the Residual Layer, $\otimes$ denotes the element-wise multiplication. We also use the attention residual learning method proposed in [10] to prevent the Masking Block from removing useful features.

Masking Block bases on the U-net structure proposed in [11], which is a famous structure to localize small medical objects. This block consists of one contracting path (encoder) and one expansive path (decoder) as shown in Figure 2b. The use of the Masking Block is the main difference of our attention modules compared to other [8], [10]. Besides, it is noteworthy that the Masking Block has a varied number of pooling and upsampling layers depending on the spatial feature size of the residual input unit. The Masking Block can play the role of activations in many segmentation architectures instead of only Unet-like architecture as in our method.

TABLE I: The detailed Residual Masking Network configuration (RL: Residual Layer, MB: Masking Block).

Layer name	Ouput size	Detail
Conv1	$64 \times 112 \times 112$	7×7, stride 2
MaxPooling	$64 \times 56 \times 56$	3×3, stride 2
Resmasking Block 1	$64 \times 56 \times 56$	RL 1, MB 1
Resmasking Block 2	$128 \times 28 \times 28$	RL 2, MB 2
Resmasking Block 3	$256 \times 14 \times 14$	RL 3, MB 3
Resmasking Block 4	$512 \times 7 \times 7$	RL 4, MB 4
Average pooling	$512 \times 1 \times 1$	
FC, Softmax	7	

C. Ensemble Method

In a competitive environment, it is difficult to avoid the impact of ensemble methods on booming accuracy. To demonstrate the ability to combine the Residual Masking Network with other CNNs, we use a simple non-weighted sum average ensemble to fuse the prediction results of 7 different CNNs. The model searching is based on the terms of validation accuracy as in [34]. The procedure of generating ensemble results is described at our GitHub.

IV. Experimental results

A. Dataset

The experiments were conducted on one published dataset and one private dataset, which is going to be public shortly. The first well-known dataset is FER2013 [22], which was introduced during the ICML 2013 Challenges in Representation Learning. This dataset, as shown in Figure 3a, contains a total of 35887 grey-scale (48x48) images. There is a total of 28709 images used for training images, 3859 for validation and 3589 for testing.

(a) FER2013 dataset

(b) VEMO dataset

Fig. 3: Example images of two datasets.

Each image is collected by Google image search API and labelled as one of the seven categories, including anger, disgust, fear, happiness, sadness, surprise and neutral. This dataset is widely used for evaluating deep learning-based FER methods. However, the dataset contains several invalid samples (e.g. non-face images or images with faces cropped incorrectly) and the image distribution among emotion categories is not equal. There are more than 6000 images showing happiness (H) while the number of images containing disgust (D) is just approximately 500.

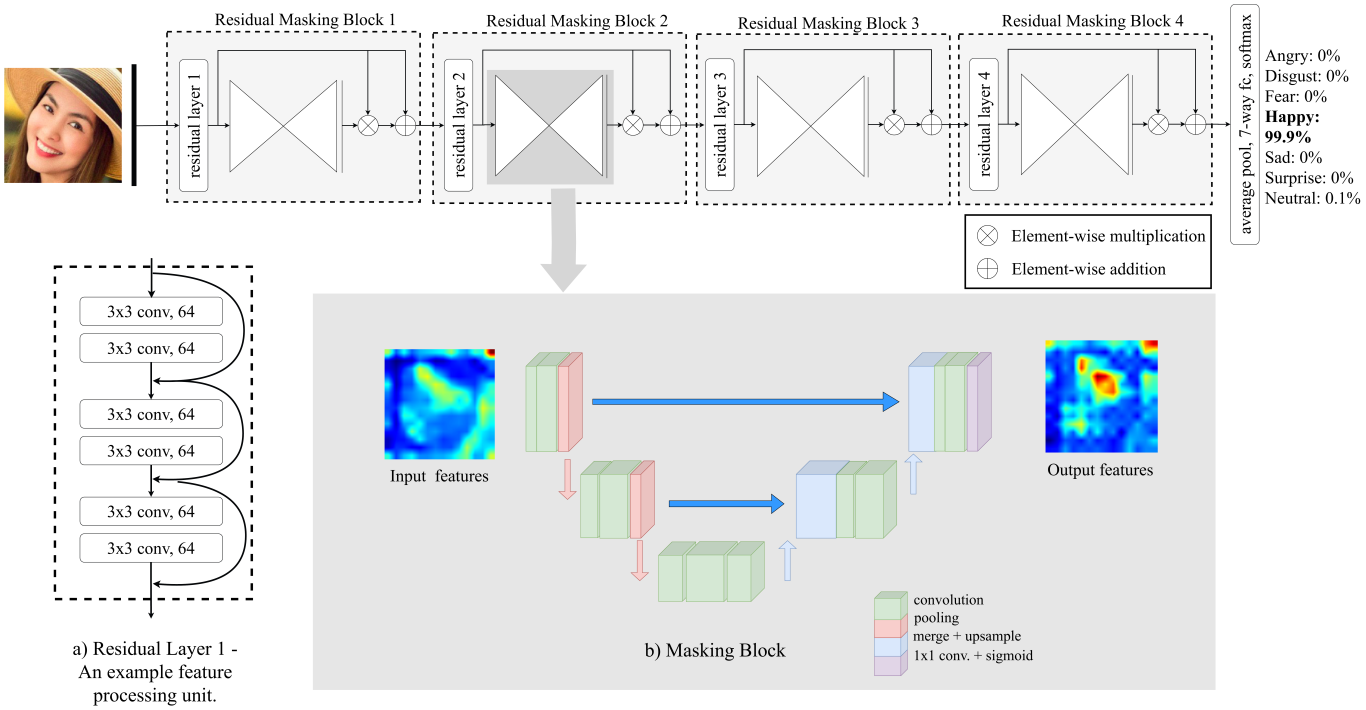


Fig. 2: The overview of Residual Masking Network.

The second dataset is Vietnam Emotion (VEMO2020). This dataset contains 36470 images (in multi-resolution) that are separated into two parts. The first part contains 6470 colored photos from Youtube, which are applied [35] to detect face region and then select five frames per second for each video; and from Google Image, Flickr (Vietnamese people only). The emotion of each image is chosen by the voting of a group, including ten members (each member from 18 to 23 years old). The second part includes 30000 images that were labelled by the professional in emotion labelling [20].

The examples and the distribution among facial expression categories in training/validation/testing set of this dataset are shown in Table II and Figure 3b.

TABLE II: The statistics of VEMO dataset.

	Angry	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Train	4,176	1,333	1,739	4,948	4,601	3,867	4,896
Val	860	281	378	1,054	993	839	1,069
Test	876	289	361	1,064	943	807	1,096

B. Experimental Setup

The original training images are scaled up to 224×224 and converted to RGB before the training process to adapt with ImageNet pre-trained models. Besides, training images are augmented to prevent over-fitting. The augmentation methods include left-right flipping and rotating in the range of $[-30, 30]$.

Each experiment lasts for a maximum of 50 epochs, and stop when validation accuracy is not improved more than eight steps. The batch size of 48 and the initial learning rate of 0.0001, scheduler reduces learning rate ten times if validation accuracy not increased in two continuous epochs.

The momentum is set to 0.9 and weight decay to 0.001. The evaluation metric for classification task is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. Experiments from different networks are also conducted using the same setting environment such as hyperparameters, preprocessing, augmentation, as well as evaluation metrics.

Experiments are conducted using Pytorch[2] on GTX 1080Ti. The detail of the experiments, reports, as well as the inference or testing code, is also presented in our repository. On the other hand, a laptop CPU core I7-8750H, VGA GTX 1050Ti, RAM 16GB is used for testing processing time in the real application. With this infrastructure, the proposed network can process 100 frames per second; each frame contains a single face. With this result, we can guarantee the real-time application.

C. Evaluation and Analysis

The proposed method has been tested on several aspects and produced positive results.

For the public dataset FER2013, firstly, we choose a number of well-known and powerful classification networks (such as Resnet151 [26], Densenet121 [36], Cbam_resnet50 [8], and Bam_resnet50 [9]) to evaluate our method (train/test in the same environment and dataset). Table III shows the comparison

[2]http://pytorch.org

TABLE III: Performance evaluation of well-known classification networks and our method on FER2013.

Networks	Parameters x 10^6	Accuracy (%)
VGG19 [37]	139.5	70.80
ResAttNet56 [10]	29.0	72.63
Densenet121 [36]	6.9	73.16
Resnet152 [26]	58.1	73.22
Cbam_resnet50 [8]	28.5	73.39
Our ResMaskingNet	**142.9**	**74.14**

TABLE IV: Performance evaluation (Accuracy) of reported methods without ensemble (WE) and ensemble (E) on FER2013.

Networks	E (%)	WE (%)
Human Accuracy [22]	-	65 $\pm$5
DL-SVM [38]	-	71.16
CNN-SIFT [32]	-	73.4
CNNs and BOVW + local SVM [33]	75.42	-
Ensemble 8 CNNs [34]	75.2	-
Our ResMaskingNet	**76.82**	**74.14**

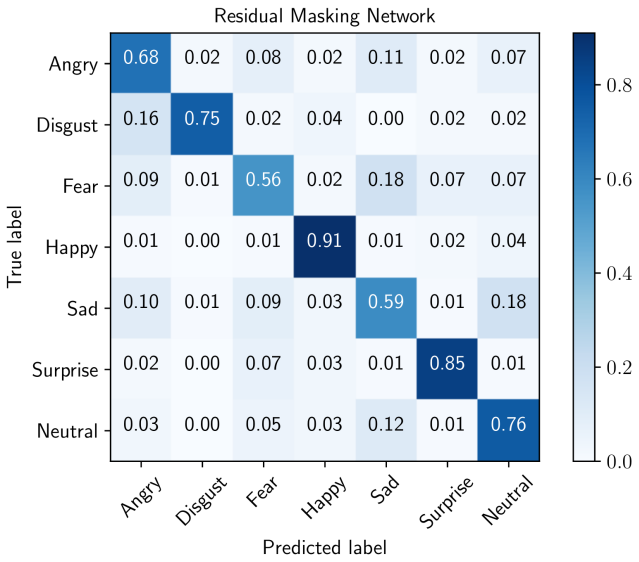
between them in the accuracy as well as the number of parameters. As presented in this table, although our Residual Masking Network has the highest number of parameters, it outperforms recent well-known deep learning-based classification networks.

Besides, the detailed evaluation with the SOTA networks on the FER2013 (ensemble /without ensemble mode) is presented in Table IV. With the ensemble mode, our Residual Masking Network ensembled with 6 CNNs obtained the highest result of 76.82%, outperformed all ensemble-based methods on FER2013 by 1%. On the other hand, our single Residual Masking Network received the highest mark of 74.14%, more top than the runner-up(73.14%) of the CNN-SIFT [32]. The CNN-SIFT approach aggregates feature from both deep learning and traditional SIFT algorithm.

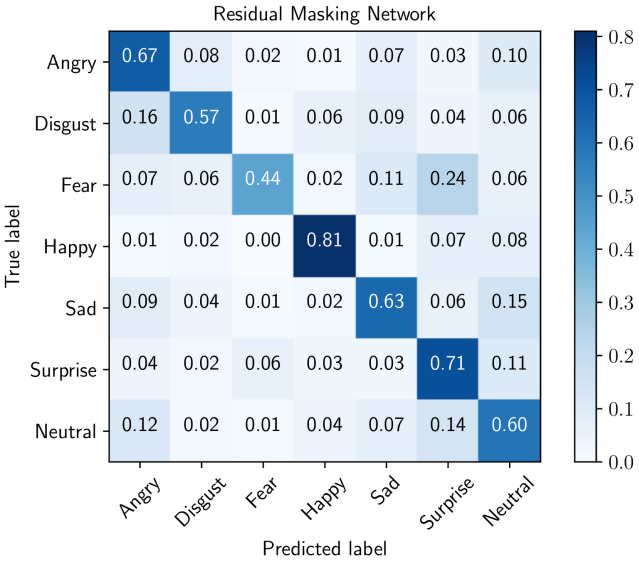
For the VEMO dataset, we conducted four experiments to compare the performance of our Residual Masking Network with three other classification networks as following: Resnet18, Resnet34, and ResAttNet56 (see Table V).

TABLE V: Performance evaluation of three classification networks and our Residual Masking Network on the VEMO dataset.

Networks	Accuracy (%)
ResAttNet56 [10]	60.82
Resnet18	63.94
Resnet34	64.84
Our ResMaskingNet	**65.94**



(a) FER2013 dataset



(b) VEMO dataset

Fig. 4: Quantitative results in form of confusion matrix on two testing sets

Besides the accuracy comparison, the Residual Masking Network's performance is also evaluated using confusion matrix as shown in Figure 4 where each row of the matrix represents instances of a predicted class while each column shows illustrations of the true label. The matrix values indicate that the network performs well on almost facial expressions even though the FER2013 or VEMO training dataset is unbalanced across classes. The Happy and Surprise class obtained the high scores of 0.91/0.81 and 0.85/0.71 while the performance of Fear and Sad are just 0.56/0.44 and 0.59/0.63, which are the lowest scores among evaluation scores. Example of the detection errors can be found in Figure 5.

We can see the ability of Masking Block in boosting the

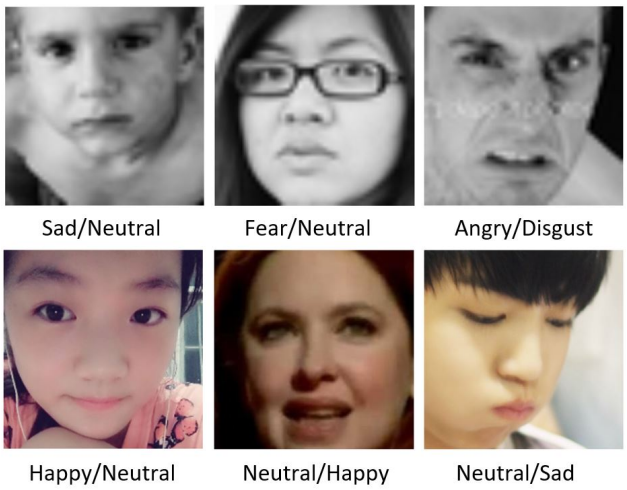


Fig. 5: Examples of wrong recognition on FER2013 and VEMO dataset (Ground truth / Predict)

accuracy of the Resnet 34 model on both FER2013 and VEMO datasets. As shown in Figure 1, where the activations before (column 3) and after (column 4) the 3^{rd} Masking block are visualized, the heat map seems to prefer the eyes, nose, and mouth to other facial regions.

These results also appear to be in line with reality for a person's emotional recognition; it is a quite subjective process [39]. Most of the wrong predictions come from wrongly labelling or the unclear emotion of the subjects, see Figure 5. About the wrongly labelling, Barsoum et al. [39] tried to fix it, and they also produce the FER+ dataset, which is less noisy than the original one. A research result presented in [22] shows that the people's ability to recognize other human's emotions is just above the average. As presented in Table IV, the Human accuracy (accuracy given by a Human estimation) is only 65% with an error of 5%. The most recognizable emotions via facial expression are happy and surprise, whereas the complex emotions like sadness or fear are difficult to distinguish. Despite not playing a decisive role, data imbalance makes it a bit difficult for emotional recognition algorithms.

Rare emotions like fear or disgust are harder to identify with either humans or machines. Emotion is still a challenging topic when people themselves are still very much confused about their feeling. This is also an exciting feature of this research direction.

V. Conclusion

This paper put forward a system for facial expression recognition, in which the main contribution is a novel Masking Idea that is implemented in the Residual Masking Network. This Residual Masking Network contains several Masking Blocks which are applied across Residual Layers to improve the network's attention ability on important information. Experimental results showed that the proposed methods possess higher accuracy than the well-known classification systems as well as the current state-of-the-art reported results on the FER2013 dataset. The focus on the future improvement of the proposed method is checking out the model generalization by evaluating it on the largest classification dataset, the ImageNet dataset. Furthermore, different network parameters, as well as model parameters reduction, will be explored to improve network performance across vision tasks such as classification and detection. We have a goal of building a complete system and conducting testing on an open rehearsal environment.

Acknowledgment

We would like to express our thanks to Mr. Pham Van Linh (phamvanlinh143@gmail.com), Dr. Nguyen Trung Kien (tkienng253@gmail.com), and Ms. Pham Huong Linh (linh.phamhuong@gmail.com) for their support in our research.

References

[1] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019.

[2] A. Mehrabian, "Communication without words," *Communication theory*, pp. 193–200, 2008.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[4] L. S. . D. Weihong, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.

[5] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in face detection and facial image analysis*, pp. 63–100, Springer, 2016.

[6] W. H. I. Tang, *Facial expression recognition for a sociable robot*. PhD thesis, Massachusetts Institute of Technology, 2007.

[7] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *International Conference on Artificial Neural Networks*, pp. 84–94, Springer, 2018.

[8] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[9] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[10] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[12] R. Li, W. Liu, L. Yang, S. Sun, W. Hu, F. Zhang, and W. Li, "Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 3954–3962, 2018.

[13] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564 – 578, 2016.

[14] Z. G. SHong X., Xu Y., "A tensor unfolding revisit," in *Computer Vision – ACCV 2016 Workshops, Lecture Notes in Computer Science*, p. 10116, 2017.

[15] H. Ali, D. Powers, X. Jia, and Y. Zhang, "Extended non-negative matrix factorization for face and facial expression recognition," *International Journal of Machine Learning and Computing*, vol. 5, pp. 142–147, 2015.

[16] D. G. . J. Lee., "Geometric feature-based facial expression recognition in image sequences using multi-class ada-boost and support vector machines," *sensors*, vol. 13, pp. 7714–7734, 2013.

[17] S. H. . A. G. . A. Routray., "A real time facial expression classification system using local binary patterns," in *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on.*, pp. 1–5, 2012.

[18] B. C. Ko., "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, p. 401, 2018.

[19] 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*, pp. 5562–5570, 2016.
[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
[21] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.
[22] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, pp. 117–124, Springer, 2013.
[23] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 472–478, 2016.
[24] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *international conference on machine learning*, pp. 2217–2225, 2016.
[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
[26] 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*, pp. 770–778, 2016.
[27] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM international conference on multimodal interaction*, pp. 553–560, 2017.
[28] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with cnn ensemble," in *2016 international conference on cyberworlds (CW)*, pp. 163–166, IEEE, 2016.
[29] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343–350, 2017.
[30] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *2015 international joint conference on neural networks (IJCNN)*, pp. 1–8, IEEE, 2015.
[31] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, IEEE, 2018.
[32] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid cnn–sift aggregator," in *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pp. 139–149, Springer, 2017.
[33] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.
[34] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint arXiv:1612.02903*, 2016.
[35] P. V. . M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, pp. 511–518, IEEE, 2001.
[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[38] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.
[39] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 279–283, 2016.