Region Adaptive Self-Attention for an Accurate Facial Emotion Recognition

Seongmin Lee*, Jeonghaeng Lee*, Minsik Kim, and Sanghoon Lee†
Yonsei University, Seoul, Korea
E-mail: {lseong721, leedoright, minsik.kim, slee}@yonsei.ac.kr, Tel/Fax: +82-2-2123-7734

***Abstract*—Recognizing facial emotion is important in human communication, especially non-verbal communication. Despite the recent advancements in deep learning, facial emotion recognition has not achieved high performance compared to other classification tasks. Motivated by the mechanism of human visual perception, in which humans recognize the facial emotion by combining the informative facial regions (i.e., eyebrows, eyes, nose, and mouth) with different weights, we propose a novel facial emotion recognition network. To effectively train the informative facial regions, we introduce *adaptive patch extraction* and *region adaptive self-attention* schemes. The adaptive patch extraction initially decides the informative facial region based on the human facial perception. Then, based on the decided informative facial regions, attention weights between regions are estimated from the region adaptive self-attention scheme. Finally, by combining the features of facial regions with attention weights, the proposed network accurately recognizes facial emotion. The experimental results show that the proposed network effectively focuses on the informative region of the human face. Furthermore, through the comparison of facial emotion recognition accuracy, it is verified that the proposed network remarkably outperforms the state-of-the-art methods.**

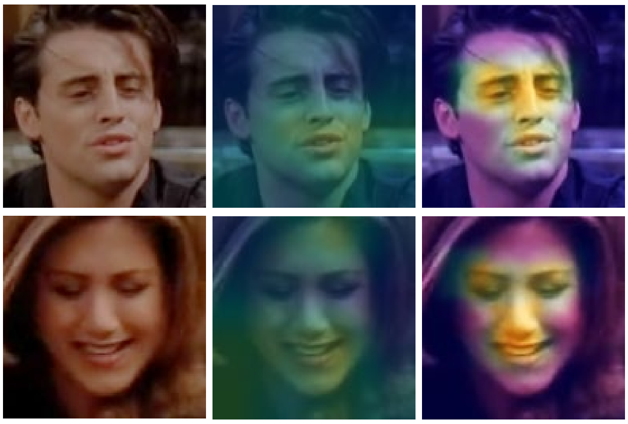


Fig. 1. Visualization of the input image and attention maps: (a) input image, (b) attention map from the conventional grid self-attention of [14], and (c) attention map from the proposed region adaptive self-attention.

I. Introduction

A facial expression can be interpreted as an informative indicator that visually hints at a change in a person's emotional state. These characteristics can be effectively used as a means for non-verbal communication between people in various fields such as psychiatry [1], lie detection [2], human-computer interface (HCI) [3] as well as online-meeting, which has recently heightened in demand. When a person's face information is given, a human can recognize the emotion in a glance based on the social experience. However, instructing an AI model to automatically infer human emotions in the analogous context as humans, there still exists a large void. Therefore, the topic of facial expression recognition (FER) from visual content having diverse domains (e.g., image, video, and recently 3D virtual reality) has been actively studied [4], [5], [6], [7].

With the recent increase in computing power and the advent of various deep learning model architectures, fields such as image classification or facial analysis (e.g., face landmark detection, face parsing, and age/gender prediction) achieved very high accuracy. However, despite the fact that the problem of recognizing emotions from human faces can be regarded as a kind of classification problem, the performance remains at an accuracy of about 60%. It is very low performance compare to the conventional image classification deep network, which achieves almost 90% accuracy. Such insufficient performance hinders the room where artificial intelligence-based emotion recognition technology can be utilized throughout the industry.

To solve this problem, we focus on the human visual perception mechanism of recognizing an facial image. According to the research on the human visual system, it is discovered that humans perceive facial emotion by concentrating on a specific facial component area, such as the eyes and mouth [8], [9], [10]. The upper part of the forehead, which usually occupies most of the facial region, has little effect on emotion perception. Also, each region's importance varies depending on the emotion types [11]. Thus, semantically important regions need to be considered with different weights. In addition, by combining the expressions of facial components, humans finally judge facial emotion [12], [13]. Inspired by this human visual perception mechanism, we propose a novel facial emotion recognition network by effectively extracting and aggregating facial components features.

To enable the network to extract features effectively from the facial component regions, we propose the *adaptive facial patch extraction* and *region adaptive self-attention* schemes. Through the adaptive patch extraction, informative facial regions, such as the eyes and mouth, are extracted. It enables the network to more focus on the informative facial regions than the

[*] These authors contributed equally to this work
[†] Corresponding author

background area. In addition, we introduce region adaptive self-attention to assign different weights on every facial patch according to the emotional condition. Figure 1 shows the attention maps from the region adaptive self-attention compared to the conventional self-attention scheme. It shows that the attention map from region adaptive self-attention is more highlighted in the facial component regions, and it is highly correlated to the human visual perception. In this paper, inspired by the human visual perception mechanism above, we propose a novel facial emotion recognition network by introducing the adaptive patch extraction and region adaptive self-attention schemes.

II. Related Work

A. Facial Emotion Recognition

Before deep learning was actively used for facial expression recognition tasks, studies using hand-crafted features based on image processing have been mainstream. To obtain expression information, methods such as geometric features indicating the shape and location of faces[15], [16], Local Binary Pattern for person-independent recognition[17], and scale-invariant feature transform (SIFT) were utilized[18], [19].

In the recent decade, deep neural models, especially convolutional neural networks, have been actively used as a feature extractor that maps some data to a high-level space [20]. Various large-scale FER target databases have been released[21], [22], [23], and a number of methods showing high-performance for facial expression recognition were proposed[24], [25], [26]. Vo *et al.*[27] proposed pyramid with super-resolution method for in-the-wild images and achieved very high accuracy. Additionally, a number of studies have been conducted to learn networks more effectively using a custom loss function. Farzaneh *et al.*[28] adaptively selected significant features, and Fard *et al.*[29] used a method to correlate feature vectors according to similar classes to enhance discriminative ability of a model. Since these methods handled all the facial parts with the same weight, the informative facial regions cannot be emphasized in the network. In this work, by adopting the region adaptive self-attention scheme, the network can effectively learn the different weights depending on the facial regions.

B. Self-Attention of Transformer Architecture

From the seminal work of Dosovitskiy *et al.* [30], the Transformer architecture using the self-attention scheme becomes popular in various tasks. The Transformer was first successful in many natural language processing (NLP) tasks. Due to the outstanding performance of the Transformer, many attempts have been made to apply its concept to computer vision. Recently, the Vision Transformer (ViT) [14] was proposed which applies the self-attention scheme in the image classification. The ViT divides the input image into several grid patches and regards each patch as a token in NLP. From the grid patches, image features are extracted and attention between patches is computed. The ViT achieves excellent performance compared to the convolutional neural network while requiring considerably fewer computational costs.

Various studies based on ViT have been conducted [31], [32], [33]. However, these methods use grid patches to compute self-attention. Due to the grid patch, facial semantic information for emotion recognition may be distorted. As we extract the image patches using the adaptive patch extraction scheme based on human visual perception, facial semantics can be well preserved. Therefore, the proposed network can learn more accurate attention between semantic facial regions for facial emotion recognition.

III. Proposed Method

The overall framework of the proposed facial emotion recognition network using region adaptive self-attention is described in Fig. 2. The proposed network comprises three parts: *adaptive patch extractor, region adaptive self-attention*, and *patch aggregation layer*.

A. Adaptive Patch Extractor

Humans does not perceive all facial regions with the same weight [8], [10]. Especially when perceiving emotions from the face, humans focus on some specific regions, such as eyebrows, eyes, and mouth, rather than the forehead and jaw. Inspired by this, we firstly extract informative facial regions from the input facial image. However, selecting the informative regions from the entire face is crucial, which is directly related to performance. Based on the fact that the facial landmark is an important factor to perceive emotion accurately [34], we define informative facial region using facial landmark. Thus, the adaptive patch extractor extracts a set of image patches to represent informative facial regions using each landmark as a center point. Additionally, using only the facial patches can result in loss of facial global context information, which is also important for recognizing facial characteristics. To add the global context, we down-sample the input image to the patch size and append to the set of image patches. The patch for global context is appended to the end of the image patch set. Thus, the adaptive patch extractor, $PE(\cdot) : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{(N+1)\times H_p\times W_p\times C}$, is formulated as:

$$I_p = PE(I), \tag{1}$$

where I is input image, I_p is a set of image patches, H, W, C are the height, width, and channel of the input image, N is the number of landmark, and H_p, W_p are the height and width of the patches. For the landmark detection, we used off-the-shelf facial landmark detector [35]. In our experiments, we used 51 landmarks ($N = 51$), excluding those on the jaw, among the landmarks defined in [35]. As we use the global context patch, the total number of used patches in our experiments is 52.

B. Region Adaptive Self-Attention

With the advent of the Vision Transformer (ViT) [14], self-attention becomes popular in computer vision. To enable the network to learn the facial informative regions effectively, we adopt the self-attention mechanism of the ViT.

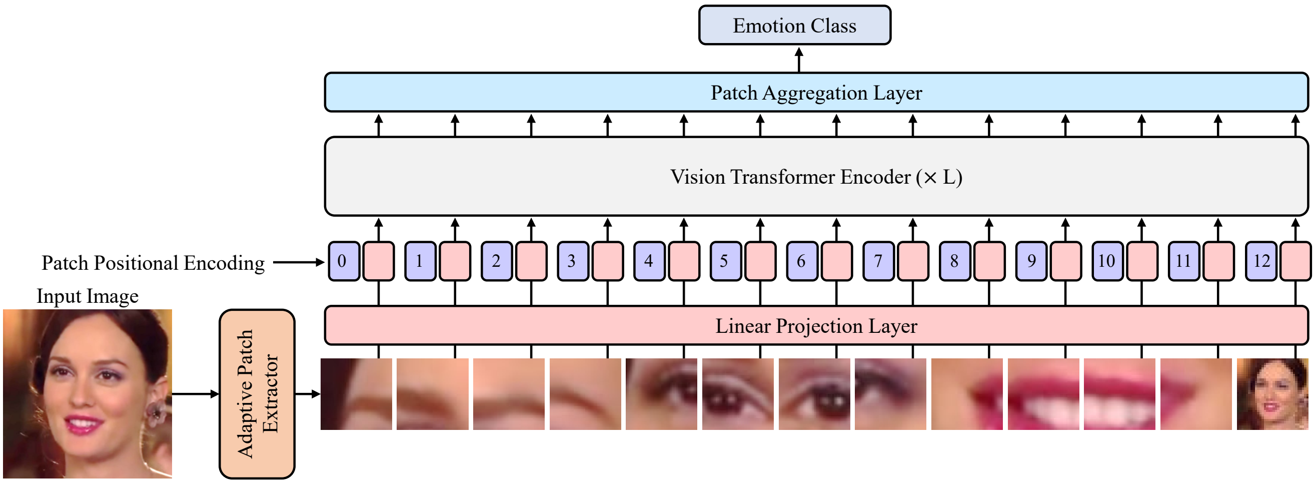


Fig. 2. An overall framework of the proposed facial emotion recognition network using region adaptive self attention.

Instead of directly using the patches as encoder input, we map the patches into D dimensional embedding space through the linear projection layer. As we construct the facial patches based on the landmark, the order of patches follows the landmark order and each patch indicates a specific facial region. Thus, position information should be added to the patches to effectively train the unique statistics of each patch. Following the previous work [14], we use standard learnable one dimensional positional encoding. The embedding vectors of the image patch through the linear projection layer is represented as:

$$\mathbf{x}_0^n = \mathrm{flatten}(I_p^n)\mathbf{E} + \mathbf{E}_{pos}, \tag{2}$$

where n is the index of the patch, $\mathbf{E} \in \mathbb{R}^{(H_p \times W_p \times C) \times D}$ is weight of linear projection layer, and $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$ is patch positional encoding. In our experiments, we set the dimension of embedding space D to 64.

The sequence of embedding vectors to which the patch positional encoding is added serves as input to the ViT encoder. The ViT encoder comprises L layers of multi-head self-attention and feed-forward blocks. We use the standard self-attention using query, key, and value.

$$Q^n, K^n, V^n = \mathbf{x}_0^n \mathbf{W}_{\mathrm{SA}}, \tag{3}$$

$$SA\left(\mathbf{x}_0^n\right) = \mathrm{softmax}\left(\frac{Q^n K^n}{\sqrt{D}}\right) V^n, \tag{4}$$

$$MSA\left(\mathbf{x}_0^n\right) = \left[SA_1\left(\mathbf{x}_0^n\right); \cdots; SA_k\left(\mathbf{x}_0^n\right)\right] \mathbf{W}_{\mathrm{MSA}}, \tag{5}$$

where $SA(\cdot)$ is self-attention operator, $\mathbf{W}_{\mathrm{SA}} \in \mathbb{R}^{D \times 3D}$ is weight to embed query Q, key K, and value V of dimension D, and $MSA(\cdot)$ is multi-head self-attention operator with projection matrix $\mathbf{W}_{\mathrm{MSA}} \in \mathbb{R}^{k \cdot D \times D}$. We set number of multi-head k to 16 in our experiments.

The output of l^{th} Transformer encoder layer is formulated as:

$$\hat{\mathbf{x}}_l^n = \mathrm{MSA}(\mathrm{LN}(\mathbf{x}_{\mathrm{l}-1}^{\mathrm{n}})) + \mathbf{x}_{l-1}^n, \tag{6}$$

$$\mathbf{x}_l^n = \mathrm{FF}(\mathrm{LN}(\hat{\mathbf{x}}_l^n)) + \hat{\mathbf{x}}_l^n, \tag{7}$$

where $\mathrm{LM}(\cdot)$ is layer normalization operator, and $\mathrm{FF}(\cdot)$ is feed-forward network consists of two linear layers with a Gaussian error linear unit (GELU) [36] non-linear activation function. The number of Transformer encoder layer L is set to 6 in our experiments.

C. Patch Aggregation Layer

Instead of using the class token in [14], we aggregate the embedded feature vectors of all facial patches. Through the patch aggregation layer, feature vectors of all patches are simply aggregated using global average pooling. Then the emotion recognition is performed by projecting the embedded feature space to the emotion class space. Thus, the patch aggregation and emotion recognition is formulated as:

$$\mathbf{x} = \frac{1}{N+1} \sum_{n=1}^{N+1} \mathbf{x}_L^n, \tag{8}$$

$$\mathbf{y} = \mathrm{softmax}(\mathbf{x}\mathbf{W}), \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{D \times D_e}$ is projection matrix from embedding space to the D_e dimensional emotion label space and $\mathbf{y}$ is predicted emotion.

To train the proposed network, we use conventional binary cross entropy loss function. Thus, the loss function is defined as:

$$loss = \mathrm{BCE}(\mathbf{y}, \hat{\mathbf{y}}) \tag{10}$$

where $\hat{\mathbf{y}}$ is ground-truth of emotion class and $\mathrm{BCE}(\cdot,\cdot)$ is the binary cross entropy function.

IV. Experiments

A. Dataset and Implementation Details

We used two public facial emotion datasets to train the proposed network: AffectNet [21], RAF [22], [23], and CAER-S [37]. AffectNet contains 450,000 real-world images annotated with eight discrete facial emotions categories and is collected from the internet by querying expression-related keywords.

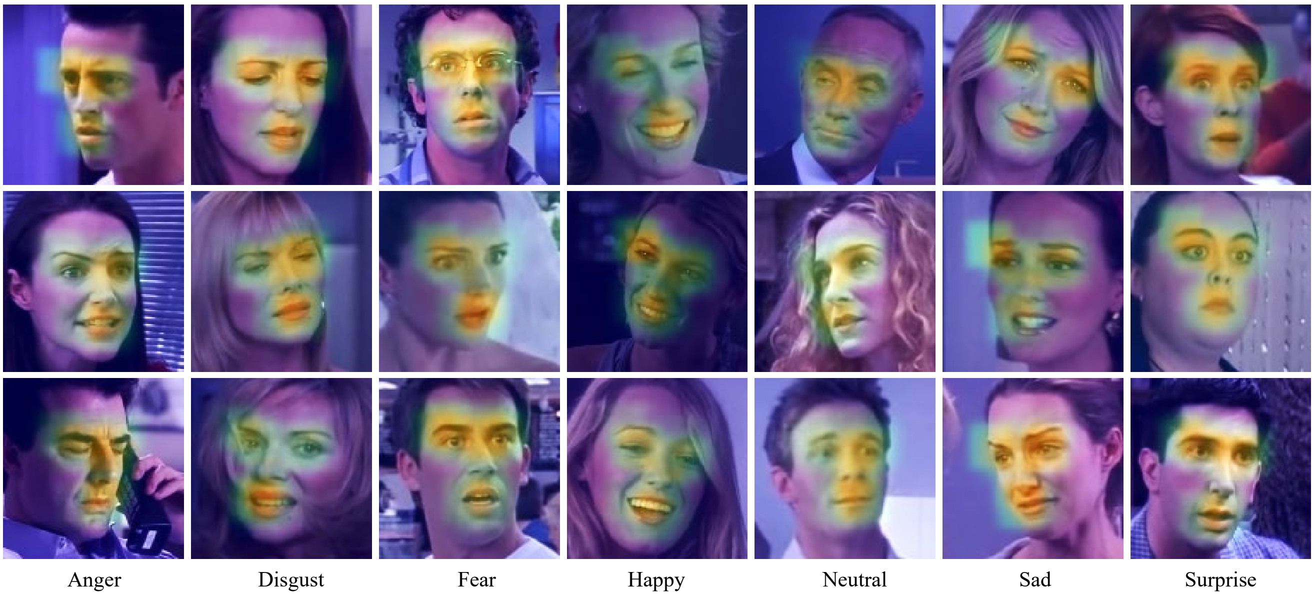


Fig. 3. Visualization of the attention map from the predicted output to the input image space on CAER-S dataset.

TABLE I
EMOTION RECOGNITION ACCURACY ACCORDING TO THE PATCHES EXTRACTION ON CAER-S DATASET

Method	Grid	Adaptive
Accuracy	0.605	0.758

TABLE II
EMOTION RECOGNITION ACCURACY ACCORDING TO THE PATCH SIZE ON CAER-S DATASET

Patch size	8 × 8	16 × 16	32 × 32	64 × 64
Accuracy	0.663	0.691	0.758	0.755

TABLE III
EMOTION RECOGNITION ACCURACY ACCORDING TO THE NUMBER OF PATCHES ON CAER-S DATASET

# of patches	51 (w/o Global)	52 (w Global)	69 (w Global)
Accuracy	0.706	0.758	0.757

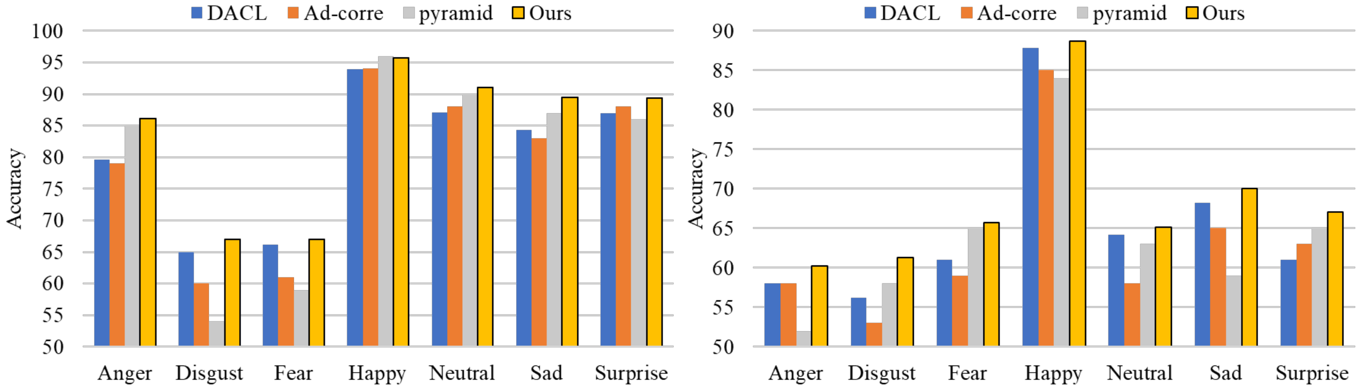
RAF includes 30,000 images obtained from real-world. Each image is annotated with one of seven emotion categories. CAER-S includes 70,000 images extracted from videos of 79 TV shows. Each image is annotated with seven emotion categories. To train the network, we select seven emotion categories (i.e., anger, disgust, fear, happy, neutral, sad, and surprise) that overlapped in both datasets. Monte-Carlo cross-validation with 10 repetitions was performed to ensure the network performance.

The proposed network used in all experiments was implemented using Pytorch library. For optimization, we used Adam optimizer [38]. The default hyperparameters were set to $\beta_1 = 0.9$ and $\beta_2 = 0.009$. The learning rate was set to 0.001 and decreased by 0.95 for each of the 20 epochs. We resized the input facial image to 256×256 ($H = W = 256$) and extracted facial patches with the size of 32 ($H_p = W_p = 32$).

B. Evaluation on the Region Adaptive Self-Attention

To analyze the performance of using the adaptive region, we visualized the attention map from the predicted emotion label to the input image space. We compare the attention map of the proposed network, which used adaptively extracted patches, with the attention map obtained using grid patches. For the grid patch extraction, we divide the input image into a set of patches of 32×32 sizes, which is the same size as the adaptively extracted patch.

Figure 3 shows the attention maps of the proposed network. Compared to networks using grid patches, the attention maps of the proposed network highly focus on informative regions, such as the eyes and mouth. In particular, when a person opens his eyes or mouth widely, that part is activated largely compared to the other parts. It confirms that the proposed network recognizes facial images similar to human visual perception. In addition, to quantitatively evaluate the performance obtained by using adaptively extracted patches, we measure the emotion recognition accuracy depending on the patch extraction method. Table I summarizes the emotion recognition accuracy of using the adaptive patch extractor. It shows that the proposed adaptive region extractor enables the network to achieve higher accuracy than using grid patch. In conclusion, we find that the proposed adaptive region extraction makes the network work in a similar way to the human visual perception mechanism and allows higher accuracy to be achieved.



(a) RAF

(b) AffectNet

Fig. 4. Quantitative comparison of emotion recognition accuracy between the proposed network and other state-of-the-art methods (DACL[28], Ad-corre[29], and pyramid[27]) for each emotion category on RAF[22] and AffectNet[21] datasets.

C. Evaluation on the Patch Characteristics

Since the proposed network uses the image patches as input, network performance is largely dependent on the characteristics of the patches. Therefore, we evaluate network performance according to the patch size and number.

Table II summarizes the emotion recognition accuracy according to the patch size. We vary the patch size from 8×8 to 64×64 multiplied by a factor of 2. The results show that using the 32×32 patch has the most higher accuracy. If the patch size is too small (i.e., 8×8 patch), the cropped patches can not contain enough facial regions, resulting in lower accuracy. In contrast, if the patch size is too large (i.e., 64×64 patch), all patches have similar global facial information, so the patches are not distinct. It prevents the network from learning the unique facial regional features, resulting in lower network performance. Therefore, using the mid-size patch allows the network to achieve the best accuracy.

In addition, since the network takes a set of image patches as input, network performance may depend on the number of patches. Table III summarizes the emotion recognition accuracy according to the number of used patches. The results show that it is most inaccurate to use only patches extracted from 51 landmarks except for the jaw. Also, it is confirmed that using the global context increases the accuracy. This is because without the global context, the network cannot learn the entire facial shape. On the other hand, increasing the number of patches does not improve accuracy, as shown in the results of using 69 patches that uses all landmarks, including the jaw, for the patch extraction. It confirms that using the landmarks of eyebrows, eyes, nose, and mouth is enough to represent the facial emotional features.

D. Comparison with the State-of-the-art Methods

To evaluate the emotion recognition accuracy, we compare the proposed method with three state-of-the-art methods: DACL[28], Ad-corre[29], and pyramid[27]. Fig. 4 shows the performance on 7 category emotion labels in RAF and AffectNet datasets. In both datasets, the proposed method has outstanding accuracy in all emotion categories. The results of the RAF dataset (Fig. 4a) show that *disgust* and *fear* have lower accuracy than other emotion categories. Such results arise from the similarity of disgust and fear images in the RAF dataset. Nevertheless, the proposed method shows higher accuracy than other state-of-the-art methods in disgust and fear by achieving both 67% accuracies, because the proposed method can effectively focus on informative facial regions. In the AffectNet dataset (Fig. 4b), the proposed method achieves an accuracy of 88.7% in *happy*, which is the highest accuracy among the state-of-the-art methods. In addition, the results show that the proposed method shows stable performance by achieving more than 60% accuracy in all emotion categories. Thus, it is confirmed that the region adaptive self-attention is effective to distinguish facial emotion and the proposed method outperforms other state-of-the-art methods in all datasets.

V. Conclusion

In this paper, we have proposed a novel facial emotion recognition network with the region adaptive self-attention scheme. By introducing the self-attention scheme based on the mechanism of human visual perception, the proposed network more effectively utilizes facial informative regions, which is crucial in emotion recognition. The quantitative and qualitative experimental results demonstrated that the proposed network achieves significant improvements over the state-of-the-art methods. We hope that this work will be applied to performance improvement in various facial emotion-related applications.

Acknowledgment

This work was supported by the Industrial Technology Innovation Program (No. 20012603, Development of Emotional Cognitive and Sympathetic AI Service Technology for Remote (Non-face-to-face) Learning and Industrial Sites) funded By the Ministry of Trade, Industry and Energy (MOTIE, Korea)

References

[1] L. Collin, J. Bindra, M. Raju, C. Gillberg, and H. Minnis, "Facial emotion recognition in child psychiatry: a systematic review," *Research in developmental disabilities*, vol. 34, no. 5, pp. 1505–1520, 2013.

[2] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, vol. 1, no. 2, p. 5, 2009.

[3] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of face recognition*. Springer, 2005, pp. 247–275.

[4] J. Kang, S. Lee, and S. Lee, "Competitive learning of facial fitting and synthesis using uv energy," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 5, pp. 2858–2873, 2021.

[5] S. Heo, H. Song, J. Kang, and S. Lee, "High-quality single image 3d facial shape reconstruction via robust albedo estimation," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1428–1432.

[6] X. Liu, L. Jin, X. Han, J. Lu, J. You, and L. Kong, "Identity-aware facial expression recognition in compressed video," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7508–7514.

[7] H.-S. Cha and C.-H. Im, "Performance enhancement of facial electromyogram-based facial-expression recognition for social virtual reality applications using linear discriminant analysis adaptation," *Virtual Reality*, vol. 26, no. 1, pp. 385–398, 2022.

[8] M. Xu, Y. Ren, and Z. Wang, "Learning to predict saliency on face images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3907–3915.

[9] J. Kang and S. Lee, "A greedy pursuit approach for fitting 3d facial expression models," *IEEE Access*, vol. 8, pp. 192 682–192 692, 2020.

[10] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[11] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Saliency-based framework for facial expression recognition," *Frontiers of Computer Science*, vol. 13, no. 1, pp. 183–198, 2019.

[12] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face." *Journal of personality and social psychology*, vol. 37, no. 11, p. 2049, 1979.

[13] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PloS one*, vol. 12, no. 5, p. e0177239, 2017.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2011, vol. 1.

[16] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 2005, pp. 76–76.

[17] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion recognition from arbitrary view facial images," in *European Conference on Computer Vision*. Springer, 2010, pp. 490–503.

[20] C. Huang, "Combining convolutional neural networks for emotion recognition," in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*. IEEE, 2017, pp. 1–4.

[21] 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.

[22] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[23] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[24] A. V. Savchenko, "Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2359–2366.

[25] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[26] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International conference on multimodal interaction*, 2019, pp. 562–566.

[27] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[28] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2402–2411.

[29] A. P. Fard and M. H. Mahoor, "Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26 756–26 768, 2022.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[32] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[34] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 2783–2788.

[35] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[36] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[37] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.