Facial Expression Recognition using Visual Transformer with Histogram of Oriented Gradients

Jieun Kim, Ju o Kim, Seungwan Je , Deokwoo Lee*
Department of Computer Engineering, Keimyung University, Daegu,South Korea

Abstract

Emotions play an important role in our life as a response to our interactions with others, decisions, and so on. Among various emotional signals, facial expression is one of the most powerful and natural means for humans to convey their emotions and intentions, and it has the advantage of easily obtaining information using only a camera, so facial expression-based emotional research is being actively conducted. Facial expression recognition(FER) have been studied by classifying them into seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and normal. Before the appearance of deep learning, handcrafted feature extractors and simple classifiers such as SVM, Adaboost was used to extracted Facial emotion. With the advent of deep learning, it is now possible to extract facial expression without using feature extractors. Despite its excellent performance in FER research, it is still challenging task due to external factors such as occlusion, illumination, and pose, and similarity problems between different facial expressions. In this paper, we propose a method of training through a ResNet [1] and Visual Transformer [2] called FViT and using Histogram of Oriented Gradients(HOGs) [3] data to solve the similarity problem between facial expressions.

1.Introduction

Emotions play an important role in everyday life as a response to our interactions with others, decision-making, and what happens around us. Based on this, emotion recognition research is essential to develop human-like intelligence systems [4]. Over the years, many psychologists and engineers have conducted research on analyzing facial expressions, voice, text, gestures, and physiological signals to understand and classify emotions [5]. Facial expressions, one of the various emotional signals, are one of the most powerful and natural means for humans to convey their emotions and intentions, and have advantage of accessibility by simply obtain information using only cameras. In addition, psychologists' report shows that facial expressions make up 55% of the effectiveness of communication messages, while language and voice make up 7% and 38%, respectively. [6] Therefore, facial expression-based emotion recognition research can improve human-computer interaction system performance. In addition, many studies have been conducted with practical importance in social robots, medical care, driver fatigue monitoring, and many other human-computer interaction systems [7]. Facial expression recognition(FER) is categorized into two stages, feature extraction and feature classification, based on seven basic emotions (anger, disgust, fear, happiness, sadness, and surprise) defined by Ekman and Friesen and including neutral expression [5] [8] [9]. As a traditional method before the advent of deep learning, hand-crafted descriptor and classifiers are used to

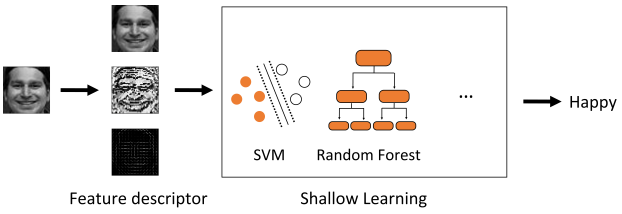

Figure 1. *Traditional method of FER*

classify facial expressions.Representative handcrafted descriptor for FER are HOGs [3], LBP [10], and NMF [11]. Extracted face features through these descriptor, classifiers such as SVM [12] and Adaboost [13] is used to classify facial expressions. Handcrafted methods shows relatively high accuracy on In-Lab dataset, but still shows low accuracy on Wild dataset [14]. With the advent of deep learning, CNN allows us to learn without using feature descriptor, and shows relatively high accuracy compared to FER using traditional methods. Furthermore, various CNN-based models have been proposed, demonstrating superior performance in FER. Despite such extensive research activities, there are still challenges due to external factors such as occlusion, illumination, and pose. In addition, in the case of expressions, there is a problem that the similarity between different expressions of the same person (intra-class) is high, and the similarity between the same expressions of different human faces (inter-class) is measured low. FER based on attention mechanism [15] has been actively conducted to reduce the accuracy deviation arisen from external factors, and the results have shown significant improvement.

Nevertheless, the problem of similarity between different expressions and within the same expression has not been solved. To solve this problem, this paper proposes a method called FViT which is to learn data that has undergone feature extraction preprocessing of images through HOGs [3], a traditional FER feature

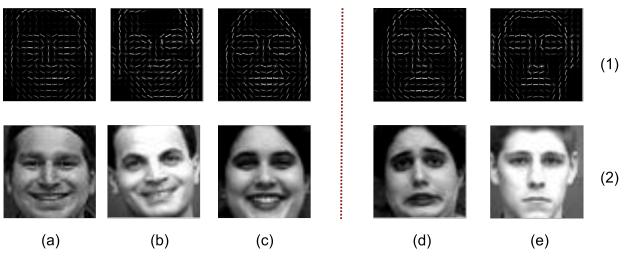

Figure 2. *Sample examples of facial data used for the experiments, (1) HOG images, (2) Original images from CK+ datasets*

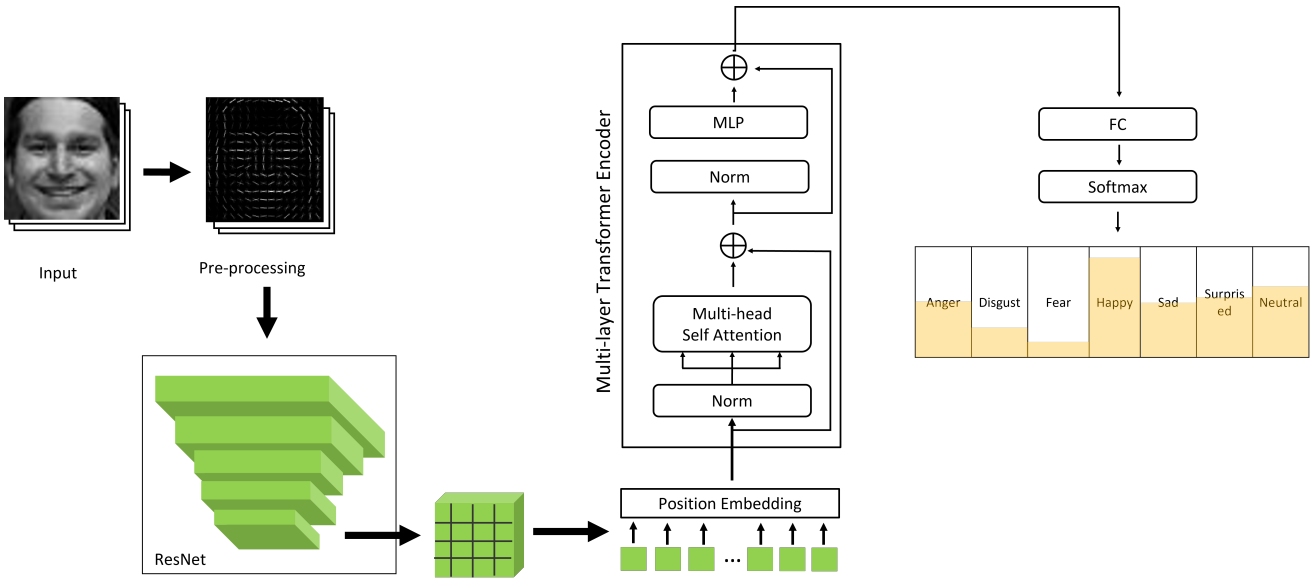

Figure 3. *Proposed model*

extractor, through model combined with CNN-based ResNet [1] and Visual Transformer [2].

2.Proposed Methods

In case of FER, there are various manual functions such as LBP [10] and HOGs [3] developed for facial feature extraction [14].However recent FER studies use deep features extracted through CNN to distinguish facial expressions.Despite the development of FER through deep learning, many problems exist. Typically, a CNN-based model requires a large amount of training data to avoid overfitting, but since the existing facial expression data does not have enough data, learning through the existing CNN-based model may cause an overfitting problem. In this paper, we propose an image preprocessing using HOGs feature descriptor [3] which can solve the overfitting problem and generalize only facial expression data in FER image. Figure 2 shows (1) HOGs image and (2)RGB image from CK+ datasets. More specifically the case of (a), (b), (c) is image from 'happy' class ,and (d),(e) is image from 'sad' class. Existing FER methods are mostly trained by CNN-based model using RGB images such as (2). In the case of facial expression data, although it is a different facial expression of the same person as the images (c) and (d), learning through RGB images has high similarity, which makes classifying facial expression difficult. To solve this problem, we proposes using HOGs images as learning data, which further highlights facial expression information and remove the unique facial information of individuals. In this section, we ex-

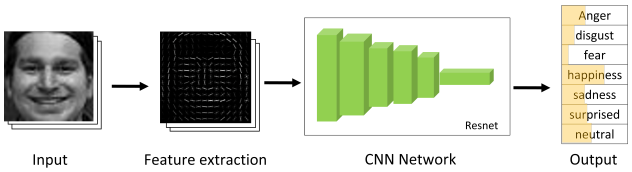

Figure 4. *Experimental results using HOGs+ResNet*

Table 1. Comparison accuracy between with HOGs and RGB CK+ data in ResNet architecture

Methods	Accuracy(%)
ResNet18	97.79
ResNet34	98.18
ResNet18 + HOGs	**99.79**
ResNet34 + HOGs	99.39

periments using the resnet [1] model, which showed high accuracy in image classification. Figure 4 shows brief structure of resnet model we experimented. Table 1 shows accuracy comparison between using HOGs and RGB images as training data. The table 1 shows that in Resnet18 model training with HOGs images increased 2% accuracy higher than the training with RGB images. Furthermore, in ResNet34 using HOGs images increased 1.21% accuracy higher than using RGB image. Through this experiments, it was confirmed that using HOGs data showed higher accuracy in classifying facial expressions than using RGB data. In the field of image classification, many people attempted to apply transformer into computer vision tasks.Visual Transformer is the first method to apply a transformer to image classification research, and the transformer-based approach shows better performance than the CNN-based method when it is pre-trained on large-scale data [14]. ViT [2] needs large datasets to generalize well on computer vision tasks. Afterwards, some work to combine the convolutional layer with the transformer was proposed, which further improved the performance of the transformer. Inspired by this work, we propose a method using HOGs [3] im-

ages as training data for ViT [2] models. Figure 3 shows the architecture of the FViT model we proposed. First we extracted HOGs [3] image from local area of training data. We calculate by using 24 orientations and (16,16) size of cell. By HOGs image, we extracted facial expression features. Result of the first experiment, ResNet18 showed the best accuracy, we used fourth layer of ResNet18 for extracting feature map from HOGs image. In this paper, we used ViT [2] model for transformer network. The extracted feature map is divided into 16 patches and flattened into vector then embedded through linear operation. After the class token and position embedding is added the input data for ViT [2] is complete.The embedded patch matrix is normalized by layer normalization, and training is accomplished through the process of repeating the Transformer Encoder composed of layer normalization, multi-head self-attention and multi-layer perceptron.

3.Experimental Results

Table 2. Performance Comparison with state-of-art methods on CK+ dataset

Methods	Accuracy(%)
FER-IK [16]	97.59
LBP+HOG [17]	98.3
ALSG [18]	93.08
FMPN [19]	98.06
FViT(w/o HOG)	96.97
FViT	**99.39**

CK+ Datasets : The Extended Cohn Kanada (CK+) [20] database is the most used laboratory control data in the field of FER. CK+ contains 593 video sequences on 123 topics. The duration of the sequence is 10 to 60 frames, showing the transition from expressionless to highest expression. In CK+, seven basic expressions of anger, contempt, disgust, fear, happiness, sadness, and surprise are used as labels based on the Facial Action Coding System (FACS). In this paper, a total of 981 frames were extracted and used in the experiment, and the experiment was randomly divided into 800 training images and 181 test images. We used 6 labels of data (anger, disgust, fear, happiness, sadness, and surprise) with neutral expression for experiment. In this experiment, the extracted CK+ data were resized to 224×224. All experiments were conducted at 100epochs, and Adam optimizer was used. In addition, the ReduceLronPlateau function was used to prevent falling into the local minima by the gradient descent method. Table 2 shows performance comparison with state-of-art methods on CK+ dataset. The methods used in comparison is as follows, knowledge-augmented image-based FER model(FER-IK) [16], LBP and HOG features with SVM(LBP+HOG) [17],auxiliary label space graphs(ALSG) [18],facial motion prior networks(FMPN) [19],the proposed methods FViT with and without(w/o) HOGs. Through the experiment, it can be confirmed that, when experimenting with the FViT model, the accuracy increase 2.422 % when using HOGs image compared to RGB image used by training data. In addition,through the the experimental results shown in Table 2, FViT showed highest performance on CK+ datasets.

4.Conclusion

In this paper, we proposed a new methods for FER, to solve the problem of similarity between different expressions and within the same expressions. To solve this problem, we propose a method of training through FViT which is conducted by ResNet [1] and Visual Transformer [2] with using HOGs feature descriptor [3] to pre-processing training data. Based on the experimental results, we found out that using feature descriptor like HOGs in pre-processing training data increases accuracy in FER. In addition, FViT showed 99.39% in CK+ datasets, which is the highest performance in comparison with other FER methods. Also from experimental results, we expects using feature descriptor in data augmentation will increases the model accuracy in FER. Therefore, in future work, based on this research, we intend to find a way to further increase the accuracy of the FER model through an experiment to apply feature descriptor to the data augmentation.

References

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

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

[3] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*], **1**, 886–893, Ieee (2005).

[4] Dave, C. and Khare, M., "Emotion detection in conversation using class weights," in [*2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*], 231–236, IEEE (2021).

[5] Li, S. and Deng, W., "Deep facial expression recognition: A survey," *IEEE transactions on affective computing* (2020).

[6] Mehrabian, A., "Communication without words," in [*Communication theory*], 193–200, Routledge (2017).

[7] Lajevardi, S. M. and Hussain, Z. M., "Automatic facial expression recognition: feature extraction and selection," *Signal, Image and video processing* **6**(1), 159–169 (2012).

[8] Ekman, P. and Friesen, W. V., "Constants across cultures in the face and emotion.," *Journal of personality and social psychology* **17**(2), 124 (1971).

[9] Ekman, P., "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique.," (1994).

[10] Shan, C., Gong, S., and McOwan, P. W., "Robust facial expression recognition using local binary patterns," in [*IEEE International Conference on Image Processing 2005*], **2**, II–370, IEEE (2005).

[11] Buciu, I. and Pitas, I., "Application of non-negative and local non negative matrix factorization to facial expression recognition," in [*Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*], **1**, 288–291, IEEE (2004).

[12] Dino, H. I. and Abdulrazzaq, M. B., "Facial expression classification based on svm, knn and mlp classifiers," in [*2019 International Conference on Advanced Science and Engineering (ICOASE)*], 70–75, IEEE (2019).

[13] Wang, Y., Ai, H., Wu, B., and Huang, C., "Real time facial expression recognition with adaboost," in [*Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*], **3**, 926–929, IEEE (2004).

[14] Ma, F., Sun, B., and Li, S., "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing* (2021).

[15] Li, Y., Zeng, J., Shan, S., and Chen, X., "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing* **28**(5), 2439–2450 (2018).

[16] Cui, Z., Song, T., Wang, Y., and Ji, Q., "Knowledge augmented deep neural networks for joint facial expression and action unit recognition," *Advances in Neural Information Processing Systems* **33**, 14338–14349 (2020).

[17] Liu, Y., Li, Y., Ma, X., and Song, R., "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors* **17**(4), 712 (2017).

[18] Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., and 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*], 13984–13993 (2020).

[19] Chen, Y., Wang, J., Chen, S., Shi, Z., and Cai, J., "Facial motion prior networks for facial expression recognition," in [*2019 IEEE Visual Communications and Image Processing (VCIP)*], 1–4, IEEE (2019).

[20] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in [*2010 ieee computer society conference on computer vision and pattern recognition-workshops*], 94–101, IEEE (2010).

Author Biography

Jieun Kim received her Bachelor degree in computer engineering from Keimyung University, Daegu, Republic of Korea in February, 2021. Since then she has been working on her graduate research and is pursuing a master degree in the areas of computer vision, image and signal processing.

Ju O Kim received his Bachelor degree in computer engineering from Keimyung University, Daegu, Republic of Korea in February, 2020. Since then he has been working on his graduate research and is pursuing a master degree in the areas of computer vision, image and signal processing.

Seungwan Je is an undergraduate researcher in image, signal and information processing group in Keimyung University, Daegu, Republic of Korea. He has been working on pattern recognition and image processing based on deep learning and machine learning approaches. He is pursuing a bachelor degree in computer engineering

Deokwoo Lee received his BS, MS and Ph.D degrees all in electrical engineering from Kyungpook National University, Daegu, Republic of Korea and North Carolina State University, NC, USA, respectively. He worked at Washington University in St.Louis, MO, USA as a postdoctoral research associate before he joined Samsung Electronics, Republic of Korea as a senior researcher. He started to work as a faculty member at Youngsan University, Busan, Republic of Korea until February, 2018. From March 2018, he has been a faculty member in the department of computer engineering at Keimyung University, leading image, signal and information processing group. His areas of interest cover image and signal processing, computer vision and pattern recognition.