Journal of Defense Acquisition and Technology, Vol. 2, No. 1, Jun. 2020, pp. 33-37
https://doi.org/10.33530/jdaat.2020.2.1.33
eISSN : 2671-9401

Real-time Facial Expression Recognition using 3D Appearance and Geometric Network for Public Security

Byung-Gyu Kim

Abstract

Facial expression recognition is a promising technology that can be used in various fields of artificial intelligence. There are several challenges to deal this task, one of which is recognizing a facial expression in sequential input data for real-time recognition. Many types of convolutional neural networks have been proposed to consider temporal information together. Typically, there are appearance networks that utilize facial appearance information, geometric networks that use facial landmark points, and hybrid approaches that combined these two networks. In this paper, we compare various methods proposed so far and identify the best approach for each network. We also have an experiment the initial architecture of our proposed 3D appearance and geometric network based on these researches and validate that its 97.22% accuracy on CK+ dataset is comparable to state-of-the-art techniques.

Key Words: Artificial intelligence (AI), Facial expression recognition (FER), Emotion recognition, Deep learning, LBP feature, Appearance network, Geometric network, Convolutional neural network (CNN)

I. INTRODUCTION

In recent years, facial expression recognition techniques have been being more accurate due to advances of various deep learning approaches. Facial expression recognition techniques can be widely applicated in variety of fields such as medical, health care, robotics and self-driving vehicles. There are several challenges in this technique. The first is that it is highly dependent on datasets and the second is about applying its dynamic features. Most of application fields of facial expression recognition require real time responses of facial expression. Thus, there are a lot of attempts to improve facial expression recognition accuracy by extracting temporal features. In this paper, we discuss many researches for more accurate facial expression recognition and propose a new 3DAGN (3D appearance and geometric network) pipeline for detecting well both appearance and temporal features. At last, we mention about our simple experiment.

This paper is organized as follows. In section 2, related works are introduce with four detailed categorizations. Section 3 includes our proposed method. Section 4 will mention experiment results and section 5 concludes this paper.

II. RELATED WORKS

2.1 Datasets

Datasets taken in the controlled environment of the laboratory for facial expression recognition like CK+ (extended Cohn-Kanade) [1] and JAFFE(Japanese female facial expression) [2] are used for many researches. On the other hand, SFEW (static facial expression in wild) [3] dataset which contains dynamic facial images that are close to the real world is one of the challenging dataset of this task. There are also researches for analzing facial expression from unaffected datasets such as CASME (the Chinese academy of sciences micro-expression) [4] and DISFA (denver intensity of spontaneous facial action) [5] dataset. For recognizing the facial expression along time axis, the dataset should have sequences to detect temporal information. Thus, nonsequential datasets like JAFFE, SFEW are not suitable for real time facial expression recognition task. Datasets available in this type of study are typically CK+, MMI [6], Oulu-CASIA [7] and AFEW

Manuscript received June 26, 2020; Revised June 19, 2020; Accepted June 28, 2020. (ID No. jdaat-2020-00001)

*Corresponding Author: Byung-Gyu Kim, Department of IT Engineering, Sookmyung Women's University, Seoul, Korea,
E-mail: bg.kim@sookmyung.ac.kr, Tel. +82-2-2077-7293
Department of IT Engineering, Sookmyung Women's University, 100, Cheongpa-ro 47-gil, Yongsan-gu, Seoul, Korea

(acted facial expression in the wild) [8]. In this paper, we use CK+ dataset for our experiment.

2.2 Temporal Appearance Networks

Temporal appearance network means the network which extract temporal features from facial input images. It can be divided into two groups depending on how it input images into the network. The first network uses a few images as input while the second uses some features extracted through series of pre-processing processes as input. [9] introduces the 3D Inception-ResNet model which uses 10 frames of sequence as input. Fan et al. [10] use 3 frames as input of CNN-RNN model and Convolution 3D model and achieved an accuracy of winner of EmotiWi 2016. Liu et al. [11], on the other hand, uses an expression video clip as a spatio-temporal manifold formed by dense low-level features. Many types of features are used as much research about this field as. Sun et al. [12] reveal the comparison of accuracy by features. According to their study, there are 8.8% accuracy improvement when recognizing facial expression with LBP (local binary pattern) [13] features compared with grayscale images.

2.3 Temporal Geometric Networks

These networks extract temporal features using geometric information of input face images. There are several geometric based approaches like Canny edge detection and AAM (active appearance model) [14], MRASM (multi-resolution active shape model) and LK-flow [15] method. Facial landmark point is also one of the typical geometric feature. [16], [17] and [18] use facial landmarks as geometric feature of their networks. Hasani et al. [9] additionally utilize landmarks to emphasis the difference between the importance of main facial components and other parts of the face which are less expressive of facial expressions. Kim et al.[19] use not whole landmark points but landmark difference in the face area of major AU's which have the most active information for geometric network.

2.4 Temporal Hybrid Networks

The method combined section 2.2 with section 2.3 is the temporal hybrid network. [19] is a kind of static hybrid network and [20], [21] are temporal hybrid networks. Jung et al.[20] use two architectures of the deep networks. Their two networks receive an image sequence and facial landmark points as input respectively. They propose a new method for integrating two separate networks. Since temporal hybrid networks deal with temporal features from two perspective, they generally perform better than networks that use only one feature.

III. PROPOSED METHOD

Our proposed method is similar to hybrid method among many approaches mentioned in section 2. Fig. 1 shows the pipeline of proposed 3DAGN.

3.1 3D Appearance Network

As we mentioned in section 2.1, the accuracies are usually higher when inputs are given as features. Thus, our proposed method also takes an LBP feature as input for getting better performance. Figure n shows that how to make LBP features. After making an LBP feature for each image, input features are stored in an array in the form of values. To encode LBP feature is shown in Fig. 2.

Our proposed method uses 3D Inception-ResNet model [9] to extract appearance features from network. 3D Inception-ResNet model is modified version of original 2D Inception-ResNet model [22]. We take shallower network than original network. Fig. 3 is the proposed 3D appearance network structure. Due to paper limits, layers are showed in block form and detailed layer configurations follow [9]. According to [19], the error is biased in the second highest expression label. Thus, we pass the labels that have best and second prediction probability of appearance network to the geometric network to overcome the error occurred at the second most likely.

3.2 Geometric Network

Geometric Network uses a landmark image pair of peak expression and non-peak expression as input. Firstly, we

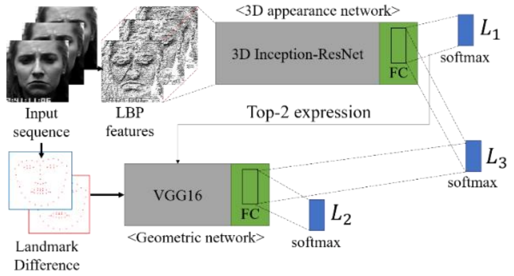

Fig. 1. Overall structure of 3DAGN.

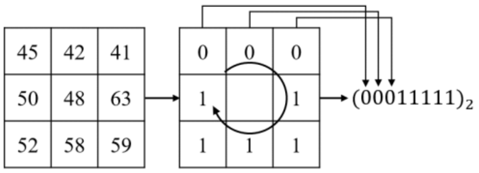

Fig. 2. Encoding an LBP pattern.

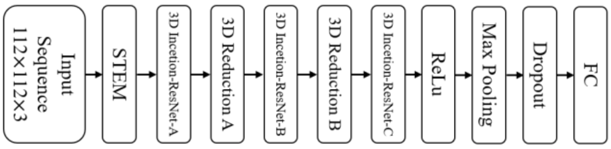

Fig. 3. Proposed network structure.

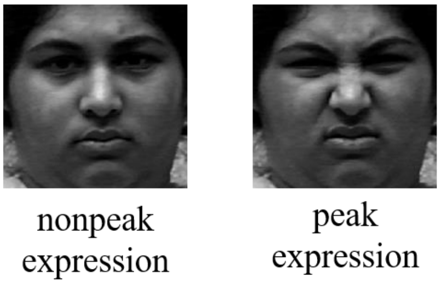

Fig. 4. Example of input expression set.

have to make expression set which consists of peak and nonpeak. Fig. 4 shows an example of expression set.

Then we can extract 68 landmark points from each of them. But not all points are used as they are. When a person makes facial expression, facial areas such as brow, eyes, nose and lips change more than other else. These parts contain a lot of information about facial expression, so we assign weights on these landmark points. After this process, landmark differences according to changes in facial expression are obtained. Then we can emphasize the change of the most expressive area. These difference images will be taken as input to the geometric network.
Geometric network is a role to find the correct expression which passed from the Top-2 expression of appearance network. This network takes VGG16 network [23] structure.

The model for each expression set is trained and stored and the model corresponding to the Top-2 from appearance network will be selected to train geometric network.

3.3. Loss Function

The loss function used to learn 3DAGN is defined as follows:

$$L_{3DAGN} = L_1 + L_2 + L_3, \tag{1}$$

where L_1, L_2 and L_3 are the loss function of 3D appearance, geometric networks and both respectively. For convenience, we call 3D appearance network, geometric network, and the integrated network by network 1, 2, and 3, respectively. Each loss function is a cross entropy loss function, which is defined as follows:

$$L_i = -\sum_{j=1}^{c} y_j^i \log(\tilde{y}_j^i), \tag{2}$$

where i is network number, y_j^1, $\tilde{y}_j^1$ and y_j^2, $\tilde{y}_j^2$ are the j-th value and j-th output value of softmax of the network 1 and 2 ground truth label respectively. Finally, c is the number of classes, L_2 has 2 classes. From last linear fully connected layer of each networks, we can get logit values. The loss function for network 3 is defined as follows:

$$L_3 = -\sum_{j=1}^{c} y_j \log(\tilde{y}_j^3), \tag{3}$$

where $\tilde{y}_j^3$ is defined as:

$$\tilde{y}_{3,j} = \sigma_s(l_{1,j} + l_{2,j}), \tag{4}$$

where $l_{1,j}$ and $l_{2,j}$ are j-th logit values of network 1 and 2. As a result, we use three loss functions in the training step and utilize only result of network 3 for test. We apply dropout method to reduce overfitting problem.

IV. EXPERIMENTS

4.1. Implementation Details

In this paper, we introduce the experiment result of the initial structure of the network corresponding to the 3D appearance network what we propose. The dataset used in this experiment is formed in the shape of 112×112×3 by cropping only face area, resizing and combining three consecutive frames from CK+ dataset. This network consists of four 3D convolution layers, three 3D max-pooling layers, one batch normalization layer and two fully connected layers with dropout. The structure of the network used in the experiment is shown in Fig. 5.

Every convolution layer has 3×3×3 size of kernel and every max pooling layer has 2×2×2 size of kernel. The active function is ReLu. At the end of the network, seven emotions are extracted by softmax function. This network is trained for about 100 epochs with Adam optimizer.

4.2. Results Comparison

Table 1 shows the experiment results on CK+ dataset of researches that performed about facial expression recognition for a sequence of several input frames including our proposed method. Comparing to other state-of-the-art works, our method achieves comparable result. Result of [20] is a combination of appearance and geometric networks, so our network will also be able to achieve outperformed results using geometric feature additionally.

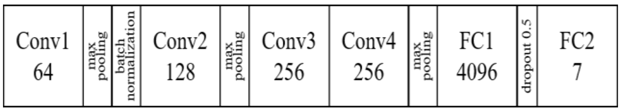

Fig. 5. Structure of 3D appearance network for experiment

Table 1. Experiment results on CK+ dataset

Methods	Accuracy(%)
2DIR with CRF [24]	93.04
3DIR [9]	93.21
STM-ExpLet [11]	94.19
DTAGN [20]	97.25
Ours 3DAGN	97.22

V. CONCLUSION

In this paper, we discuss some novel facial expression recognition approaches. Although facial expression recognition task still has difficulty that is dependent on dataset, we have confirmed that using features as input in appearance network and the hybrid method of combining appearance and geometric network are more effective for facial expression recognition. We achieve an accuracy of 97.22%, a comparable result of the state-of-the-art results by using the initial network structure which is applicable to the proposed 3D appearance network. In the future, we will organize a new network according to our proposed structure with detailed parameter tuning to get reasonable results in cross-database. Eventually, we will experiment on AFEW dataset to get outperformed result.

REFERENCES

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, United States, pp. 94-101.

[2] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Third IEEE International Conference on Automatic Face Gesture Recognition*, Nara, Japan, 1998, pp. 200-205.

[3] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark.*”* in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Barcelona, Spain, 2011, pp. 2106-2112.

[4] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, “CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces,” in *Proc. IEEE International Conference on Automatic Face Gesture Recognition*, 2013, Shanghai, China, pp. 1-7.

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

[6] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. “Web based database for facial expression analysis,” In *2005 IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, pp. 5, doi: 10.1109/ICME.2005.1521424.

[7] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietik¨ainen. “Facial expression recognition from near-infrared videos,” *IVC*, vol. 29 no 9, pp.607-619, 2011.

[8] Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. “Collecting large, richly annotated facial-expression databases from movies,*” IEEE Multimedia*, vol. 19, no. 3, 2012.

[9] B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3d convolutional neural networks,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on IEEE*, Honolulu, HI, USA, 2017 pp. 2278-2288.

[10] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” in *Proc. 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 445-450.

[11] M. Liu, S. Shan, R. Wang, and X. Chen. “Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition,*” In CVPR, 2014 IEEE Conference on*, Columbus, Ohio, USA, pp. 1749-1756. 2014.

[12] Z. Sun, Z. Hu, M. Wang, and S. Zhao, “Dictionary learning feature space via sparse representation classification for facial expression recognition,” *Artificial Intelligence Review. Jan 2019, Vol. 51 no. 1, pp. 1-18. 2019..*

[13] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, “Local binary patterns and its application to facial image analysis: A survey,”' *IEEE Transaction on System, Man, and Cybernetics. Parts C, Applications and Reviews*, vol. 41, no. 6, pp. 765-781, 2011.

[14] P. Zhao-yi, Z. Yan-hui, and Z. Yu,: “Real-time facial expression recognition based on adaptive canny operator edge detection,” *International Conference on Multimedia and Information Technology (MMIT)*, Kaifeng, pp. 154-157, 2010.

[15] B. Ahn, Y. Han, and I.S. Kweon,: “Real-time facial landmarks tracking using active shape model and LK optical flow,” *International Conference on Ubiquitous Robots and Ambient Intelligence* (URAI), Daejeon, pp. 541-543, 2012.

[16] F. Khan, “Facial expression recognition using facial landmark detection and feature extraction via neural networks,” in *Computer Vision and Pattern Recognition*, arXiv:1812.04510 [cs.CV], 2018.

[17] N. P. Gopalan, S. Bellamkonda, and V. S. Chaitanya, “Facial expression recognition using geometric landmark points and convolutional neural networks,” in *International Conference on Inventive Research in Computing Applications*, 2018.

[18] I. Tautkute, T. Trzcinski, and A. Bielski, “I know how you feel: Emotion recognition with facial landmarks,” in *Computer Vision and Pattern Recognition Workshops*, arXiv: 1805.00326 [cs.CV]

[19] J. Kim, B. Kim, P. Roy, and D. Jeong, “efficient facial expression recognition algorithm based on hierarchical deep neural network structure,” *IEEE Access*, vol. 7, pp. 41273-41285, 2019.

[20] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in *International Conference on Computer Vision*, Boston,

Massachusetts, USA, 2015, pp. 2982-2991.

[21] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193-4203, 2017.

[22] C. Szegedy, S. Ioffe, and V. Vanhoucke. "Inception-v4, inception-resnet and the impact of residual connections on learning". *arXiv preprint arXiv:1602.07261*, 2016. 2, 3

[23] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." https://arxiv.org/abs/1409.1556, 2014

[24] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Automatic Face & Gesture Recognition* (FG 2017), 2017 12th IEEE International Conference on. Washington DC, USA, 2017, pp. 790-795.

AUTHOR

Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007.

From February 2009 to February 2016, he was associate professor in the Division of Computer Science and Engineering at SunMoon University, Korea. In March 2016, he joined the Department of Information Technology (IT) Engineering at Sookmyung Women's University, Korea where he is currently a full professor.

He has published over 250 international journal and conference papers, patents in his field. His research interests include software-based image and video object segmentation for the content-based image coding, video coding techniques, 3D video signal processing, wireless multimedia sensor network, embedded multimedia communication, and intelligent information system for image signal processing. He is a senior member of IEEE and a professional member of ACM, and IEICE.