Recognition of facial expression using spatial transformation network and convolutional neural network

Jieun Kim, Eung-Joo Lee, Deokwoo Lee

Jieun Kim, Eung-Joo Lee, Deokwoo Lee, "Recognition of facial expression using spatial transformation network and convolutional neural network," Proc. SPIE 12101, Pattern Recognition and Tracking XXXIII, 121010J (27 May 2022); doi: 10.1117/12.2634030

Event: SPIE Defense + Commercial Sensing, 2022, Orlando, Florida, United States

Facial Expression Recognition Robust to Occlusion using Spatial Transformer Network with Triplet Loss Function

Jieun Kim[a], Eung-Joo Lee[b], and Deokwoo Lee[a]

[a]Department of Computer Engineering, Keimyung University, Daegu, Republic of Korea
[b]CAMCA, Dept of Radiology, MGH and Harvard Medical School, MA, Boston, USA

ABSTRACT

With the development of artificial intelligence, the field of sentiment analysis can be used in various industries such as computer-human interaction technology, personal status monitoring, criminal investigation, and entertainment. In the field of sentiment analysis, various methods such as facial expression, voice, EEG signal, and text are being studied. Among these methods, facial expression recognition is one of the approaches being actively studied because it has the advantage of being relatively easy to collect learning data and easy to apply to real life compared to other methods. Recently, research on facial expression recognition using deep learning has been actively conducted, and it shows relatively high performance. The method using deep learning has advantage of being easy to apply to a variety of data, but there is a limitation in large deviation in accuracy depending on the effect of occlusion, pose, and illumination in extracting feature points. In addition, in the case of expression recognition, similar objects such as face always exist in the data, and only some specific regions such as eyes, nose, and mouth have necessary information for learning, and remaining regions such as background and hair is considered as insignificant part of the data. Therefore learning all the features in the data not only takes a long time to learn, but also uses computing resources inefficiently.Therefore, we propose a convolutional neural network algorithm combined with a spatial transformation network which helps facial expression recognition by focusing on a specific part of the face.

Keywords: Facial expression recognition, sentimental analysis, Transformer, CNN, Deep learning

1. INTRODUCTION

Facial expressions are one of the most powerful, natural, and universal signals that convey human emotional states and intentions.[1] In the field of facial expression recognition research, many studies have been conducted due to practical importance such as human-computer interaction, driver support system, education, and entertainment.[2] Nevertheless, wild dataset in facial expression recognition has limitations in classification due to fluctuations in classes and similarities between classes due to diversity of head poses, lighting, occlusion, and personal attributes.[3] Research in the field of facial expression recognition has been mainly studied in a way of classifying through data pre-processing and learning. In the past, studies on classification using feature descriptors and classifiers have been actively conducted as learning methods, but since the announcement of Alexnet in 2012, research using CNN-based end-to-end learning has been actively conducted and shows relatively high performance.[4] The method using the feature descriptor and classifier extracts the required object from the data. Extracts the feature point vector of the extracted object using a feature descriptor such as HOG (Histogram of Oriented Gradient) and SIFT (Scale Invariant Feature Transform), and then classify using classifier such as SVM (Support Vector Machine) or Adaboost. This method has problem which is , it requires a lot of computing resources because of the uses of high-dimensional feature vector and it shows a large deviation in accuracy depending on the type of data during training. Expression recognition methods using deep neural network-based research is being actively studied by CNN such as AlexNet,[4] VGGNet,[5] GoogleNet,[6] and ResNet,[7] and various deep learning methods

Further author information: (Send correspondence to Deokwoo Lee.)
Jieun Kim. : E-,ail : lilly9928@icloud.com
Eung-Joo Lee.: E-mail: elee66@mgh.harvard.edu
Deokwoo Lee.: E-mail: dwoolee@kmu.ac.kr, Telephone: +82 53 580 5268

Pattern Recognition and Tracking XXXIII, edited by Mohammad S. Alam, Vijayan K. Asari,
Proc. of SPIE Vol. 12101, 121010J · © 2022 SPIE · 0277-786X · doi: 10.1117/12.2634030

such as RNN and GAN. The method using deep learning has the advantage of being easy to apply to various data. But it requires a large amount of training data to avoid overfitting. Also, in the case of facial expression recognition data, it has bias depending on individual characteristics such as age, gender, etc., and there is a problem with clusters overlapping due to large differences within classes and high similarity between classes,[3].[8] In this paper, we proposed method which preprocess the image by random occlusion, then generate transformed data by focusing on a specific part required for learning through a spatial transformation network. Then classify facial expressions through a modified Resnet network that combines with spatial transformer network(STN) and triplet loss.

2. PROPOSED METHOD

Although the deep learning-based classification method shows relatively high performance, in the case of facial expression recognition, there is a large deviation in accuracy due to external factors such as occlusion and posture change.[3] It is important to focus on learning. To achieve this, spatial transformation network,[9] one of the attention-focusing techniques, was used to increase learning efficiency and accuracy. In addition, we tried to solve the over-fitting problem through the pre-processing the image by random occlusion and to make it robust even in the occlusion. In this section, we introduce spatial transformer networks and loss functions and the proposed architecture(Fig. 1).

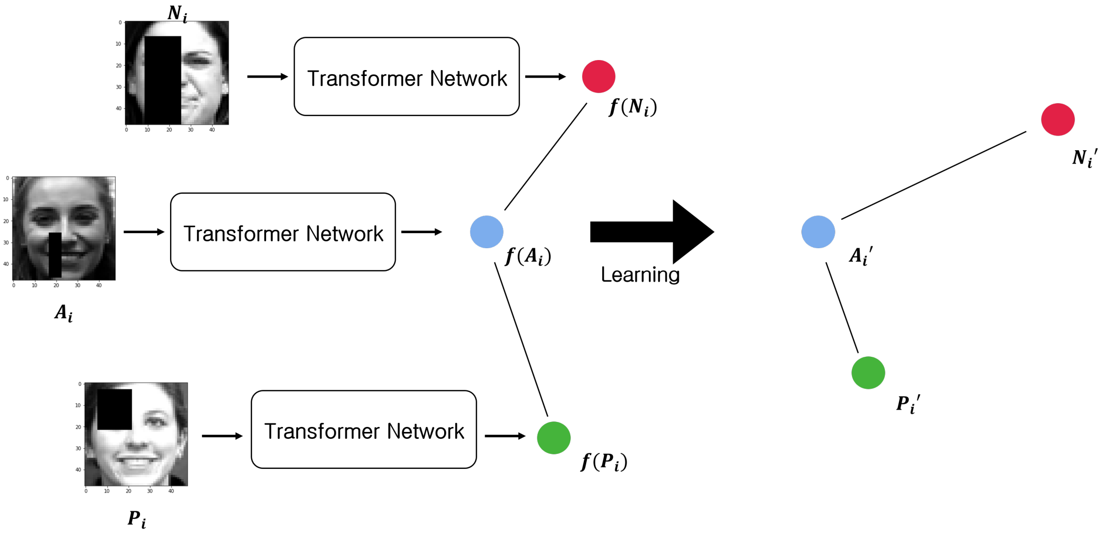

Figure 1. Overall architecture proposed in this paper for facial expression recognition.

2.1 Spatial Transformer Network

The spatial transformation network is a network that helps to intensively learn by specific part when trying to classify an image through CNN.[9] When learning deep learning-based image classification, it is a very important problem to accurately recognize even if the image is deformed (rotated, cut, occluded, etc.). The existing CNN based learning method uses the max pooling layer to solve this problem. However spatial transformation network learns by transforming a specific part required for learning ,it is more effective than using a max pooling layer.[9] Fig. 2 shows the structure of the spatial transformation network. A spatial transformation network consists of a localization network, grid generator, and sampler. The localization network learns the necessary parameters for affine transformation. The localization network receives the feature map $\mathbf{U} \subset \mathbb{R}^{H \times W \times C}$ as input, H stands for the height of the feature map, W stands for the width of the feature map, and C stands for channel. The input value feature map $\mathbf{U}$ is passed through the localization network $\mathbf{F}_{localization}$ to extract the parameters θ required for the affine transformation : $\theta = \mathbf{F}_{localization}(\mathbf{U})$. In this paper, localization network $\mathbf{F}_{localization}$ is constructed by connecting the max pooling and RELU(rectified linear unit) activation functions with two convolutional layers, respectively, and one fully connected layer. The parameter θ created by passing through

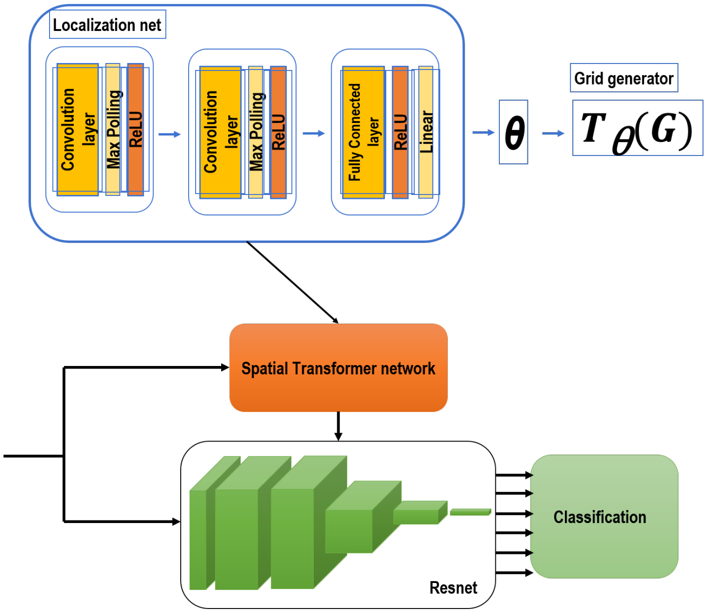

Figure 2. An illustration of Spatial Transformer Network .

is calculated the location to be sampled through the sampling grid $\mathbf{T}_\theta(\mathbf{G})$. Finally, the sampler applies the calculated sampling grid $\mathbf{T}_\theta(\mathbf{G})$ to the input feature map $\mathbf{U}$ to generate transformed output data.

2.2 Triplet Loss

Triplet loss is a loss function that first appeared in FaceNet.[10] Triplet loss calculates the loss with three values using positive and negative values based on the anchor. Triplet loss use metric learning, and it is a loss function that makes the same class set close and different class sets far away. In Eq. 2.2, the i^{th} data to be learned as an anchor A_i is received as an input. P_i receives input data in the same class as A_i, and N_i receives input data in a different class in A_i. After passing through the network, training proceeds by calculating the triplet loss as written by

$$||f(A_i) - f(P_i)||_2^2 + \alpha < ||f(A_i) - f(N_i)||_2^2 \tag{1}$$

2.3 Network Architecture

Table 1. Modified Resnet Network Diagram.

Layer type	Output size	Letter
Convolution layer1	24 ×24	7 ×7, 64, stride2
Convolution layer2	12 ×12	3 ×3, 64
Convolution layer3	6 ×6	3 ×3, 128
Convolution layer4	3 ×3	3 ×3, 256
Convolution layer5	2 ×2	3 ×3, 512
Fully connected layer	1 ×1	Average Pooling

The model proposed in this paper is largely divided into anchor, positive and negative, and then passes through a transformer network then classified into 7 expressions. The Transformer network consist of Spatial transformer network and Resnet [5]. In order to reduce the learning parameters, Resnet18 was reduced to 10 layers as shown in Table 1. When input data enters the transformer network, it passes through the spatial transformer network to generate data that transforms a specific part required for learning, and calculates the loss by passing the generated data and input data through the network configured as shown in Table 1 to be classified. This experiment was performed in the environment of Epoch 100, Learn rate 0.001, and the batch size was set to 8.

3. EXPERIMENTAL RESULTS

The experimental results are shown in Table. 2.

Table 2. Experimental results using dataset CK+

Methods	Accuracy(%)
FER-IK[11]	97.59
Nonlinear eval on SL+SSL puzzling[13]	98.23
Transformer Network (Resnet18) + Softmax	96.87
TransformerNetwork (Modified Resnet) + Triplet	**99.44**

The Extended CohnKanada (CK+) database is the most widely used laboratory control database in FER-lucey2010. CK+ contains 593 video sequences on 123 topics. Sequences vary in duration from 10 to 60 frames and show the transition from neutral to peak expression. In this video, it is classified into 7 basic expression labels (anger, contempt, disgust, fear, happiness, sadness, surprise) based on the FACS (Facial Action Coding System). In this paper, a 981 frames were extracted and used in the experiment. Here, 800 training images and 181 test images were randomly divided into experiments. In this paper, as shown in Fig. 3, pre-processing the image by random occlusion. We compared and tested data with and without random occlusion, and compared softmax loss and triplet loss. Table. 2 shows accuracy of the proposed model with other models using the CK+

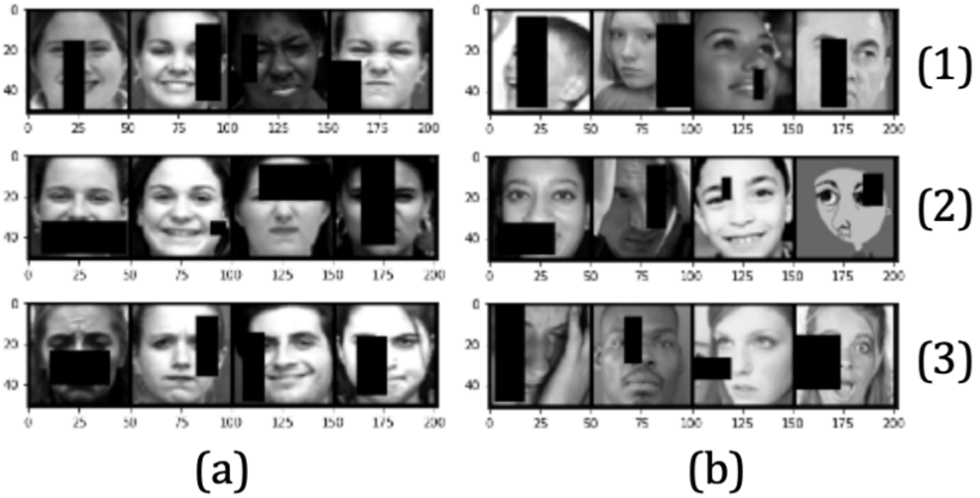

Figure 3. Sample of Datasets. (a) is CK+ database , (b) is FER2013 database. (1) is sample of anchor data , (2) is sample of positive data, (3) is sample of negative data..

database. When Softmax is used without changing the Resnet of the proposed model, it can be seen that 96.87% accuracy, which is lowest among the models. The 0.48% accuracy increased when triplet loss was used compared to modified Resnet + Softmax model. Although the ViT+SE[12] model shows a 0.35% higher accuracy than the proposed model, in this study, despite learning from data with occlusion, it shows relatively high accuracy.

4. CONCLUSION

In this paper, we proposed a model combined with STN and modified Resnet to solve the problem of accuracy deviation caused by external factors such as occlusion and posture change. In addition, the triplet loss function was applied to the clarify class classification and was compared with the softmax loss function. Proposed model(STN+modified Resnet) used in the CK+ database, 2.09% accuracy increased compared to the case of STN and resnet18 model. Also model with triplet loss increased 0.48% accuracy compared to the case of model with softmax loss. Also we could find out that using triplet loss has more clear division than when using softmax loss by visualization the data distribution.

REFERENCES

[1] Darwin, C., "The expression of the emotions in man and animals", Oxford Express(1998).

[2] Huang, Y., Chen, F., Lv. S. and Wang, X., "Facial expression recognition: A survey," Symmetry, 11(10), 1-28 (2019).

[3] Cai, J., Meng, Z., Khan, A., Li, Z., O'Reilly, J. and Tong, Y., "Island loss for learning discriminative features in facial expression recognition." arXiv:1710.03144 (2017).

[4] Krizhevsky, A., Sutskever, I. and Hinton, G., "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems 25, 1-9 (2012)

[5] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556 (2014).

[6] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition, 1-9, 2015.

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

[8] Li, S. and Weihong D., "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing(Early Access) (2020).

[9] Jaderberg, M., Karen S. and Zisserman, A., "Spatial transformer networks." Advances in neural information processing systems 28 (2015).

[10] Schroff, F., Dmitry K. and James P., "Facenet: A unified embedding for face recognition and clustering," Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 815-823 (2015).

[11] 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 (2020).

[12] Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K. and Sequier, R., "Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition ," arXiv preprint arXiv:2107.03107 (2021).

[13] Pourmirzaei, M., Gholam M. and Farzaneh E., "Using Self-Supervised Auxiliary Tasks to Improve Fine-Grained Facial Representation." arXiv preprint arXiv:2105.06421 (2021).