Transactions on Computational and Scientific Methods | Vo. 4, No. 11, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

# Enhancing Facial Expression Recognition Accuracy Through Spatial Transformation and Super-Resolution Preprocessing

**Chen Wu[1], Elliot Finch[2]**
University of Delaware, USA[1], University of Delaware, USA[2]
chenwu239@udel.edu[1], Elliot.F810@gmail.com[2]

**Abstract:** Facial expression recognition (FER) has become a critical field of study, driven by advancements in artificial intelligence and the increasing ubiquity of face recognition systems. This paper proposes an innovative FER approach that integrates image preprocessing techniques with a deep learning classification network. By employing the Spatial Transformation Network (STN), the proposed method addresses challenges such as input image size, shape variance, and background fusion. Subsequently, a super-resolution (SR) algorithm is applied to enhance image quality and preserve crucial details. These preprocessing steps feed into a traditional classification network to achieve superior recognition accuracy. Experimental results demonstrate the effectiveness of this approach, showing that it outperforms existing methods in terms of accuracy and computational efficiency. The study offers valuable insights into improving the robustness and performance of FER systems through preprocessing advancements.

**Keywords:** Facial Expression Recognition; Spatial Transformation Network; Super Resolution.

## 1. Introduction

With the rapid development of artificial intelligence and the gradual improvement of face recognition system, the research and application of facial expression recognition has become an important field of social application and scientific and technological research. Mollahosscini et al. [1] based on the GoogleNet model, a single-component network architecture is proposed. The improved GoogleNet method is used to recognize facial expressions and achieves good results. Yang Feng et al. [2] proposed a facial expression recognition method based on small-scale kernel convolution, using multi-layer small-scale convolution blocks instead of large convolution blocks to extract facial expression features to achieve expression classification.

The facial expression recognition system proposed in this paper is a classification algorithm based on a deep learning network, with an emphasis on image preprocessing at the front end of the network. First, use the Spatial Transformation Network (STN) to align the input image to improve the various problems of the input image, such as different sizes, different object shapes, and fusion with the background. Then the improved image is improved through the super-resolution (SR) algorithm to improve the overall quality and the overall information of the image. Finally, the classification result of facial expression recognition is finally obtained through the traditional classification network.

## 2. Related Work

Facial Expression Recognition (FER) has gained significant attention in recent years, benefiting from advancements in deep learning and related methodologies. Deep learning techniques have been widely explored to optimize feature extraction and image analysis. For instance, Du et al. [3] introduced HM-VGG, a multimodal deep learning framework that enhances image analysis, while Liu et al. [4] examined adversarial neural networks for semantic segmentation, offering insights into improving the robustness of feature recognition, a concept closely aligned with the preprocessing strategies in this work.

Multimodal deep learning has also shown great potential in emotion-aware systems, as demonstrated by Duan et al. [5], who proposed an intelligent user interface capable of understanding complex user emotions through multimodal integration. Such approaches underscore the importance of integrating diverse data sources, which can enhance the performance of FER systems by providing richer feature representations. Furthermore, Huang et al. [6] optimized object detection through knowledge distillation algorithms, showcasing the value of preprocessing in enhancing the overall accuracy of deep learning models, which resonates with the focus of this study on leveraging Spatial Transformation Networks (STNs) and Super-Resolution (SR) algorithms.

Luo et al. [7] addressed challenges of data sparsity and cold-start issues in recommendation systems using metric learning, highlighting the significance of data preprocessing in improving downstream performance. Similarly, Wei et al. [8] explored self-supervised graph neural networks for feature extraction in heterogeneous networks, emphasizing the importance of advanced data structuring and representation, which can inspire further advancements in FER preprocessing. Dong et al. [9] contributed to optimizing knowledge reasoning and text generation through advanced models with graph structures, illustrating transferable principles for handling complex data that could inform the development of more sophisticated FER pipelines.

In addition, adaptive strategies for resource management, such as those presented by Li et al. [10] using reinforcement learning for resource scheduling, offer valuable insights into dynamic optimization, which could influence real-time adaptive FER systems in future work. These studies collectively provide a solid foundation and inspiration for the proposed FER approach, which integrates STNs and SR algorithms to enhance the quality and robustness of preprocessing, ultimately leading to superior recognition accuracy.

## 3. Methodology

### 3.1 Spatial transformation network(STN)

The Spatial Transformation Network (STN) is mainly composed of three parts: parameter prediction (Localisation net), coordinate mapping (Grid generator), and pixel acquisition (Sampler). The main task is to complete the transformation of various spaces, expressed in formulas as follows:

$$\text{General layer}: a^{l}_{nm} = \sum\nolimits_{i=1}^{3} \sum\nolimits_{j=1}^{3} w^{l}_{nm,ij} a^{l-1}_{ij}$$

### 3.2 Super Resolution Algorithm (SR)

Super-resolution algorithm refers to the process of using software or hardware to increase the resolution of the original image, and obtaining a high-resolution image through a series of low-resolution images. The super-resolution algorithm essentially uses the known image information to predict the required pixels. In order to obtain more accurate prediction results, the prediction model of this model is much more complicated than traditional algorithms. Generally, there are multiple convolutional layers and activation layers, which use image information of a large area around the target pixel, including thousands of model parameters.

The EDSR used in this algorithm uses Generative Adversarial NetWork (GAN) to solve the problem of super-resolution. On the basis of the original ResNet network, adjustments and optimizations were made, and some unnecessary batch normalization (BN) layers in the residual structure were removed. The location of the BN layer saves 40% of the memory usage of the model during training. Therefore, a larger-scale network with better performance can be constructed with limited computing resources.

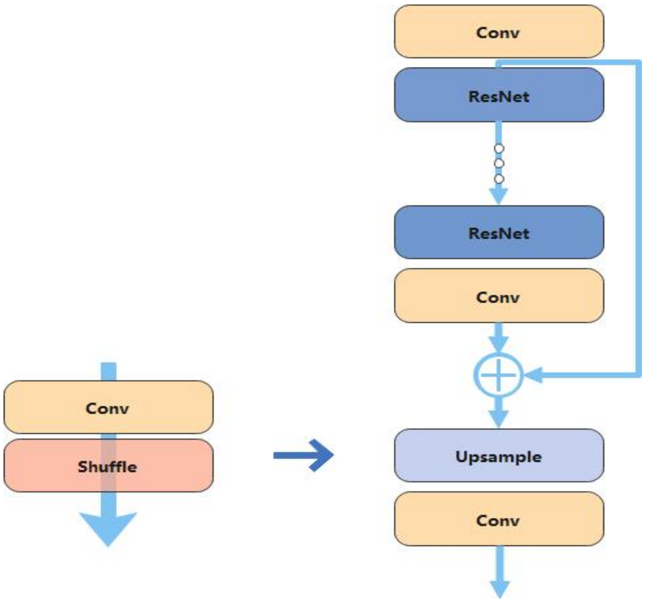


**Figure 1.** Upsample (left) and ESDR (right) structure

In the training process, the L1 norm style loss function is used to optimize the supervision network. The number of residual blocks is set to B=16, and the number of features is F=128. When training, train the low-multiple upsampling model first, and then use the parameters obtained by training the low-multiple upsampling model to initialize the high-multiple upsampling model, so that it can Reduce the training time of high-multiple up-sampling models, and at the same time get better training results.

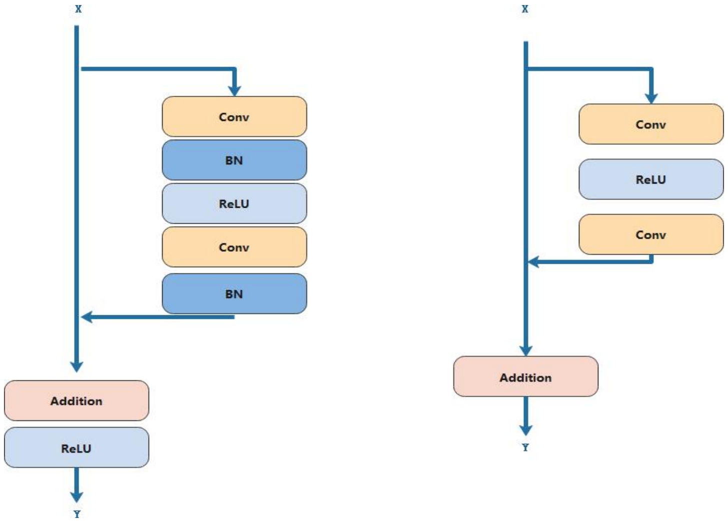


**Figure 2.** ResNet(right) and SR(left) structure

### 3.3 Algorithm specific process

In order to improve the recognition accuracy and classification robustness of the entire network, this paper proposes to improve the performance from the source through image preprocessing and reduce the amount of calculation. First, use the spatial transformation network (STN) to align the input images. Since most of the current data sets have inconsistent image sizes and the face position of each image is also different, we first need to align. The network structure of STN is as follows:

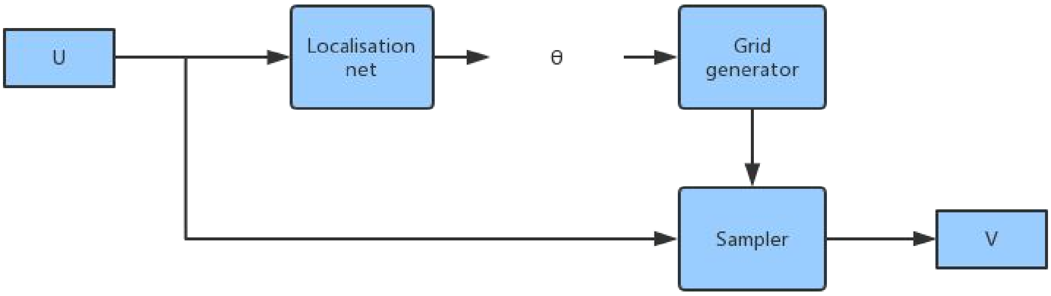


**Figure 3.** STN structure diagram

Secondly, use SR technology to expand the pixels of the original image to obtain relatively more effective information, and then extract the corresponding features through the feature extraction network. Finally, the extracted features are input to the classifier to obtain the final expression classification. The flow chart of the entire algorithm scheme is as follows:

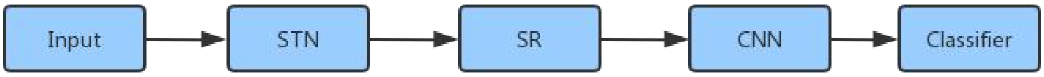


**Figure 4.** Algorithm flow

## 4. Results and discussion

### 4.1 Experimental data

The data sets used in the experiment are CK+ and FER2013, and Table 1 describes the distribution of the number of pictures in the training set and the test set. The CK+ dataset contains a total of 123 objects and 593 image sequences, of which only 327 have emoticons. The FER2013 dataset contains 35886 facial expressions, including 28708 test images (Training), 3586 public verification (Public Test) and non-public verification (Private Test) each, and each image is fixed to48×48 grayscale. The image is composed of seven expressions, correspon ding to the number labels 0-6.

**Table 1.** Quantity distribution of CK+ and FER2013

| | CK+ | | FER2013 | |
|---|---|---|---|---|
| Emotion | train | test | train | test |
| anger | 45 | 45 | 2054 | 271 |
| disgust | 59 | 89 | 107 | 15 |
| fear | 25 | 25 | 510 | 78 |
| happiness | 69 | 69 | 7335 | 895 |
| neutral | - | - | 9030 | 1102 |
| sadness | 28 | 28 | 3047 | 384 |
| surprise | 83 | 83 | 3173 | 402 |
| contempt | 18 | 18 | 115 | 13 |
| Total | 327 | 327 | 25371 | 3160 |

### 4.2 Analysis of results

First, the accuracy of the STN network is verified. In order to see the conversion effect more intuitively, it is verified on the MINIST data set. The result is shown in Figure 6, it can be seen that the STN network can obtain higher accuracy under the premise of reducing the amount of calculation, and the recognition accuracy of the network after adding the STN module can reach more than 99%. At the same time, the results in Figure 5 show that the network can achieve image space conversion, can reduce image deflection, so that the image can be extracted more easily.

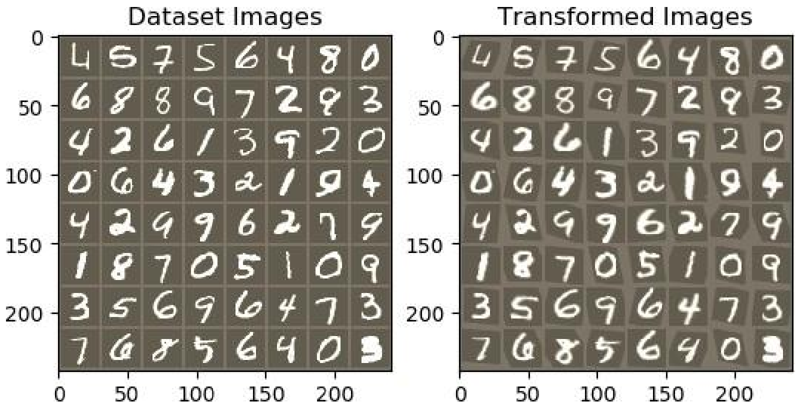
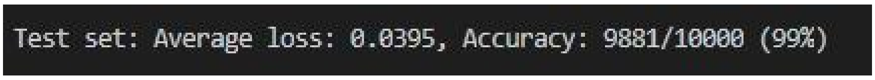


**Figure 5.** Test results of STN on MINIST



```
Test set: Average loss: 0.0395, Accuracy: 9881/10000 (99%)
```

**Figure 6.** Testing accuracy and loss of STN against MINIST

In order to verify the effectiveness of the algorithm proposed in this paper, a comparative test of multiple modules was carried out. The input is 48×48 RGB images. This model uses the ADAM optimizer, among them $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. The minibatch size is 16, and the learning rate is initially set to $10^{-4}$ update the learning rate after each $2 \times 10^5$ minibatch.

In the training process, this article uses L1 and L2 loss functions for comparison. According to the results of the comparison experiment, it can be seen that the L1 loss function is better than the L2 loss function in the network. Through analysis, the reason for this result may be because the images in this experiment contain a variety of features, corresponding to the true distribution of the relevant image composition. The underlying assumption of using L2 as the loss function is that the collected samples all belong to the same Gaussian distribution, that is, the peak value is single, but in fact, most data distributions have more than one peak value. This situation results in a low probability that the trained data set is really distributed, and the performance of the network is not good enough.

**Table 2.** Precision results of each network frame

| Network | CK+ | FER2013 |
|---|---|---|
| Base | 87.2 | 66.5 |
| Base+STN | 94.4 | 71.4 |
| Base+SR | 95.1 | 72.7 |
| Base+STN+SR | 97.8 | 73.4 |

According to the accuracy analysis of the training results of the data set, it can be seen that the accuracy of the network with the STN module is 7.2 and 4.9 percentage points higher than the accuracy of the network without the STN module in the two data sets, respectively. It can be seen that the pre-processing of the image can be Improve the performance of the network. In contrast, the accuracy of the network with the SR module is 7.9% and 6.2% higher than that of the network without the basic module, and 0.7% and 1.4% higher than the accuracy of the network with the STN module. The accuracy of the algorithm proposed in this paper is the highest in the entire experimental results, with an accuracy of 97.8 in the CK+ data set and an accuracy of 73.4 in the FER2013 data set. This shows that the algorithm proposed in this paper has high recognition performance and robustness.

## 5. Conclusion

This paper studies the performance and advantages of the spatial transformation network (STN) and the super-resolution algorithm (SR), and integrates them with the facial expression recognition feature extraction network. At present, most algorithms only aim at the modification of the feature extraction network to improve performance. The method proposed in this paper can improve the overall feature information through image preprocessing in the early stage of the recognition network, improve the robustness of the network, and reduce the calculation of the entire network. the amount. The experimental results show that the algorithm in this paper has higher performance and lower computational cost, which provides a new idea for improving the accuracy of facial expression recognition.

## References

[1] M. Tan, and Q.V. Li (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.International Conference on Machine Learning, vol.24, no.1, p. 6015-6114.

[2] Y Feng, R Liu, T Lu (2020). Facial expression recognition based on small-scale kernel convolution. computer engineering, vol.18, no.1, p.1-8.

[3] J. Du, Y. Cang, T. Zhou, J. Hu, and W. He, "Deep Learning with HM-VGG: AI Strategies for Multi-modal Image Analysis," arXiv preprint arXiv:2410.24046, 2024.

[4] H. Liu, B. Zhang, Y. Xiang, Y. Hu, A. Shen, and Y. Lin, "Adversarial Neural Networks in Medical Imaging Advancements and Challenges in Semantic Segmentation," arXiv preprint arXiv:2410.13099, 2024.

[5] S. Duan, Z. Wang, S. Wang, M. Chen, and R. Zhang, "Emotion-Aware Interaction Design in Intelligent User Interface Using Multi-Modal Deep Learning," arXiv preprint arXiv:2411.06326, 2024.

[6] G. Huang, A. Shen, Y. Hu, J. Du, J. Hu, and Y. Liang, "Optimizing YOLOv5s Object Detection through Knowledge Distillation algorithm," arXiv preprint arXiv:2410.12259, 2024.

[7] Y. Luo, R. Wang, Y. Liang, A. Liang, and W. Liu, "Metric Learning for Tag Recommendation: Tackling Data Sparsity and Cold Start Issues," arXiv preprint arXiv:2411.06374, 2024.

[8] J. Wei, Y. Liu, X. Huang, X. Zhang, W. Liu, and X. Yan, "Self-Supervised Graph Neural Networks for Enhanced Feature Extraction in Heterogeneous Information Networks," arXiv preprint arXiv:2410.17617, 2024.

[9] Y. Dong, S. Wang, H. Zheng, J. Chen, Z. Zhang, and C. Wang, "Advanced RAG Models with Graph Structures: Optimizing Complex Knowledge Reasoning and Text Generation," arXiv preprint arXiv:2411.03572, 2024.

[10] P. Li, Y. Xiao, J. Yan, X. Li, and X. Wang, "Reinforcement Learning for Adaptive Resource Scheduling in Complex System Environments," arXiv preprint arXiv:2411.05346, 2024.