Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231137

Face expression recognition based on residual network and support vector machine

Shuiyuan Liu

School of Information Science & Engineering, Lanzhou University, Gansu Province, 730000, China.

Shyliu2018@lzu.edu.cn

Abstract. In recent years, Facial expression recognition (FER) has become an outstanding research field, which content academia and industry alike. In order to alleviate the gradient disappearance in facial expression recognition caused by increasing network depth, this paper proposes joint residual network and support vector machine (SVM) to build a model. The algorithm uses a small convolutional kernel and a deep network structure. In addition, a multi-task cascaded convolutional network (MTCNN) is constructed for data pre-processing. Meanwhile, migration learning is introduced to overcome the shortcomings of insufficient data and prevent overfitting. Extensive experiments are conducted on the FER-2013 dataset. The model achieves a recognition rate of 98.6% in the test, meanwhile the accuracy of SVM's category is about 1% better than that of SoftMax. The experimental results proof that the model hereby propose outperforms different networks and previous methods. It further perceives that the method hereby propose is effective in recognizing facial expressions.

Keywords: facial expression recognition, support vector machine, multi-task cascaded convolutional network.

1. Introduction

With the rapid advancements in artificial intelligence, coupled with the increasing popularity of smart devices, face recognition technology is experiencing significant growth. The discussion surrounding this technology remains ongoing and shows no signs of slowing down. Facial expression recognition (FER) is a crucial component of face recognition technology. In recent years, it has garnered significant attention across various fields, including human-computer interaction, safety, robotics, manufacturing, automation, healthcare, communications, and driving [1]. As a result, FER has become a prominent research area in both academia and industry alike. FER involves identifying emotional and psychological changes in individuals by analyzing their facial expression states captured in still photographs or video sequences. Most traditional expression feature extraction methods (e.g., Local Binary Pattern (LBP) and Scale Invariant Feature Transform (SIFT)) rely on manually designed features [2]. Designing such features is not only challenging but also cannot guarantee their optimality. Additionally, these features cannot extract higher-order statistical features of images. There are three main directions of traditional expression recognition: overall and local recognition, deformation extraction and motion extraction, geometric features and facial features. In the holistic recognition method, both from the morphological changes about face and from the movement about face, the

DOI: 10.54254/2755-2721/21/20231137

expressive face is analyzed for the whole to identify the differences though the images with various representations. The partial recognition approaches express that variety components about faces that can be separated when the recognition processing, i.e., the weight for individual parts is different. The deformation extraction method is constructed by the deformation of different components about the face for the time expressing various results. And that movement approach is according to the doctrine that specific parts about the movement of face in response to specific expressions. This feature matrix charts a vector of features according to the form as well as the locations of the facial parts. (Nose, eyes, brow and mouth included), and this feature vector comes to represent the geometric features of the human face. Thus, researchers started to use deep learning for expression recognition. Deep neural networks have now been proven capable of mining the deep underlying distorted representational properties of data in the speech, text domains and image. Convolutional neural networks (CNNs), which are particularly effective at recognizing 2D graphics with displacement, scaling, and other forms of distortion invariance, are extensively utilized for both image identification and categorization [3-6]. The feature extraction layer of CNNs learns implicitly from the training data, avoiding the need for display feature extraction. In order to achieve successful recognition of facial emotions by deep convolutional neural networks, a significant amount of training data is essential to properly train parameters of this model. The current deep learning expression recognition algorithm mainly combines a deep learning network with expression features, due to the insufficient expression recognition databases for network parameter training. The increase in depth of CNNs greatly affects the final classification and recognition performance. In order to get rid of artificial facial expression features, deeper models are needed, as shallow networks cannot significantly improve recognition performance [4]. Nevertheless, the network depth has been found to increase with that phenomenon of gradient disappearance, which becomes more and more obvious and the accuracy decreases rapidly. He et al. proposed residual blocks to alleviate this problem [7]. The proposed deep residual network exhibits better performance than other models in multiple tasks such as object detection.

To address the phenomenon of gradient vanishing in facial expression recognition tasks caused by the increase in network depth, this paper proposes an improved algorithm that utilizes a residual network combined with a support vector machine (SVM) [8-11]. This approach does not rely on any artificial expression features and extracts deep learning features using the deep network. This study constructs a Multi-task Cascaded Convolutional Network (MTCNN) for data pre-processing [12]. ResNet-50 model network used as the core for feature acquisition, and the network is fully trained using data enhancement technology and transfer learning. Non-expression databases are used for training models and fine-tuning network parameters using expression data. After testing, the proposed model outperforms other networks and Previous methods. The experimental results have shown that the approach described in the research is working in recognizing facial expressions.

2. Methodology

2.1. Dataset description and pre-processing

The ImageNet dataset is highly utilized within the deep learning image research field. The majority of studies on image classification, localization, and detection rely on this dataset, it includes more than 14 million illustrations, representing a total of over 20,000 different profiles [5].

FER-2013 database is a dataset contributed from Kaggle face emotion recognition contest. It includes 35,887 expression-labelled pictures containing seven types of images: disgusted, angry, normal, happy, surprised, sad and fearful. Most of the database images are from the network, which contains face angles, different lighting environments, etc. and many of the images are obscured by occlusions such as hands, hair, and scarves. the CK+ expression database contains face expressions consisting of 123 people with a total of 593 sequences of expressions ranging from natural to peak. In total, it also contains eight basic expressions: angry, contemptuous, happy, sad, surprised, annoyed, scared, and neutral. In this paper, only seven of these expressions were selected for recognition in the experiments. Three to five images of each labelled expression sequence were selected as data, of which 392 (98) were selected for training

DOI: 10.54254/2755-2721/21/20231137

(testing), 56 (14) for anger, 56 (14) for disgust, 56 (14) for fear, 56 (14) for happiness, 56 (14) for sadness, 56 (14) for surprise and 56 (14) for neutrality, for a total of 392 (98) images. The experiments were performed on the CK+ database using MTCNN to crop the database for faces and then expand the data to 12 times the original data using data augmentation techniques (flip, brightness adjustment, and other operations), in addition to performing a cross-cut on each training data (test images were not used for this operation), i.e., 25600 (2268) images.

The model generalization capability was tested using the publicly available database GENKI-4K, which contains 4,000 photographs of faces in a variety of complex variations, including age, skin colour, race, pose, lighting, and environment. Of these images, 2162 photos were labelled as smiling, while 1838 photos were labelled as non-smiling. Unlike some other datasets collected in the laboratory, this dataset provides a good representation of the wide variety of challenging smiley recognition problems encountered in real life. Table 1 declines the construction of emotion dataset category.

Table 1. Construction of emotion dataset category.

Class	Training	Testing
Anger	56	14
Disgust	56	14
Fear	56	14
Happiness	56	14
Sadness	56	14
Surprise	56	14
Neutrality	56	14
Total	392	98

2.2. Proposed approach

Specifically, first, to train a machine learning algorithm, the author generally divides the raw data into a training set, validation set, and testing set. This program sets the training set, validation set, and testing set to 392, 49, and 49 examples, namely. The 48*48 size image input is sent for feature extraction to produce feature map. The data is then scaled up for 12 times the initial data using data augmentation techniques (flip, brightness adjustment, etc.). In addition to cross-cutting each training data (This operation was not used for the test images) was cross-cut. Second to build a CNN model, the author must first define three layers, the LeNet convolution pool layer, the hidden layer, and logistic regression. During the training phase of the CNN, the author must save the parameters of the model at regular intervals. The next author uses these building blocks to build CNN models. Once the CNN model is formed, it remains to be solved by an improvement algorithm, which uses the batch stochastic gradient descent algorithm (MSGD), therefore there are certain assets in the MSGD defined at the beginning, the majority of these comprise functions of cost, training validation, test mod, and parameter upgrade regulations (i.e., gradient descent). Final, the essential components of the optimization algorithm can be described and the follow step is using facial image dataset to develop models. Number of times to iterate is also set to iterate through all the samples in a batch, depending on the batch size set.

2.2.1. ResNet. To deal with the issue of gradient disappearance/inflation, this structure introduces the idea of what is named as residual block. Within this network, it used a method named as skip-connection. Skip connections connect the activation of one layer to more layers, use the way that skipping some of the intermediate layers. It creates a residual block. Residual blocks are formed by stacking these residual blocks on top of each other. The approach behind this network is that rather than having layers learn the underlying mapping, the network should adapt to the residual mapping. Thus, instead of H(x) being the initial mapping, it is better to let the network adapt mapping, as follows,

$$F(x) = H(x) - x \text{ which gives } H(x) = F(x) + x \tag{1}$$

DOI: 10.54254/2755-2721/21/20231137

The benefit of this type of skip connection being added is such that if any layer impairs the architecture's capabilities, it is skipped by normalization. This will consequently lead to training a very deep neural network without the problem of gradient disappearance/explosion. The author of the study experimented on layers 100-1000 of the CIFAR-10 dataset. There is a parallel method called "highway networks" and these use skip connections as well. Like LSTM, the skip connections use parametric gates. These gates dictate exactly the amount of information that is going to get through the skip connection. Nevertheless, this type of structure does not deliver much better precision than the ResNet structure. Figure 1 declines the Skip (Shortcut) connection.

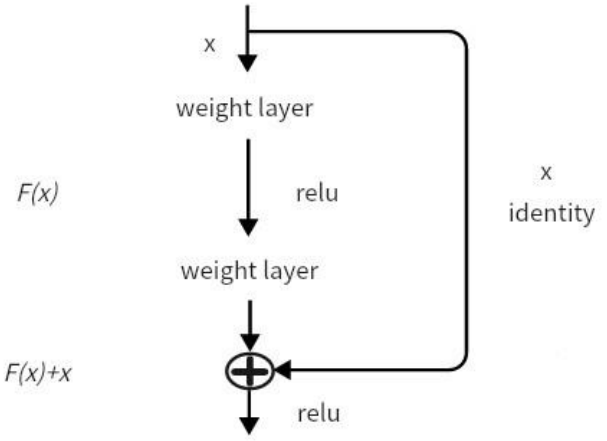

Figure 1. Skip (Shortcut) connection (Picture credit: Original).

2.2.2. SVM. A non-linear categorization problem on the input space that can be turned into a linear categorization task in some dimensional eigenspace via a non-linear transformation to enable studying a linear support vector machine in a high-dimensional eigenspace. At the time that objective and the classification decision function in a linear support vector machine learning pairwise problem just encompass the inside instance's volume, there is no need to expressly indicate the non-linear change and only the kernel function needs to be used rather than inner product. Kernel function is the inner product in the middle of specific examples obtained using a non-linear transformation. To expand, K (x, z) represent the function, which can also be called a positive definite kernel, which shows that the mapping $\emptyset(x)$ get in input and features space possesses the true mapping $\emptyset(x)$, and for any x, z in the input space, as follows:

$$K(x,z) = \phi(x) \cdot \phi(z) \tag{2}$$

For the pairwise problem in linear support vector machine learning, the inner product can be changed to the kernel function K (x, z), also final achieve path obtained after processing is the nonlinear support vector machine:

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*) \tag{3}$$

Taking the above discussion together, the non-linear support vector machine learning algorithm obtained by the author can be found below:

$$T = \{(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)\} \tag{4}$$

DOI: 10.54254/2755-2721/21/20231137

Of which

$$x_i \in R^n, y_i \in \{+1,-1\}, i = 1,2,...,N \tag{5}$$

Output: Decomposition of the hyperplane as well as the categorical decision function. Use the conforming kernel function K (x, z) and the punishment parameter C>0 as an architecture to simultaneously handle convex quadratic programming problems.

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i \tag{6}$$

$$s.t.\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{7}$$

$$0 \le \alpha_i \le C, i = 1,2,...,N \tag{8}$$

Get the best deal.

$$\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)^T \tag{9}$$

Calculation: Using a part of α^*, α_j^*, it meets the status $0 < \alpha_j^* < C$, compute:

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i K(x_i, x_j) \tag{10}$$

Classification decision function:

$$f(x) = sign(\sum_{i=1}^{N}\alpha_i^* y_i K(x, x_i) + b^*) \tag{11}$$

The SVM used a Gaussian radial basis function separator., so for these cases the function for classification decision becomes:

$$f(x) = sign(\sum_{i=1}^{N}\alpha_i^* y_i \, exp(-\frac{\|x-z\|^2}{2\sigma^2}) + b^*) \tag{12}$$

2.3. Implementation details

This experiment uses Windows10 with TensorFlow 2.6.0, python 3.9 and opencv3. And the details about hyper-parameters of this model is in Table 2.

Table 2. Hyper-parameters of this model.

Hyper-parameters	Value
Batch size	56
epoch	50
Learning-rate	0.05
Pool-size	(2,2)
n-kerns	[20,50]

DOI: 10.54254/2755-2721/21/20231137

3. Result and discussion

The present study assesses the capabilities of the model in two parts. In first part evaluates the capabilities of the suggested model by using the FER-2013 dataset, followed by the paper's analysis of the impact on model performance under different hyperparameters. Then the author competes the model suggested though this paper with other models.

3.1. The performance on the FER-2013 database dataset

The result, the author got an accuracy of 98.6 percent, which means that 7 images got the wrong Classification. In order to test the feasibility of the network, InceptionV4, VGG, and ResNet with SoftMax were compared and experimented with in this paper. After testing and training, the proposed algorithms outperform other networks and Previous methods. The experimental results have shown that the approach suggested for this area is valid in recognizing facial expressions. Table 3 is the performance for different algorithms.

Table 3. Performance for different algorithms.

Algorism	Accuracy (%)
CNN + AD	84.55
CSPL + SVM	89.89
LBP + CNN	84.40
GB + DBNs + SAE	92.46
LBP/VAR + DBN	91.40
Algorithms in this paper	98.60

The experimental results just reflect the recognition rate about ResNet with SVM algorithm outperforms traditional expression recognition algorithms. Compared with deep learning algorithms combining manual features, the algorithm avoids complex explicit feature extraction and outperforms some deep networks. The viability and validity the algorithm is proven, and it has good generalization ability in expression recognition. The algorithm has good generalization capability in expression recognition.

3.2. The influence of hyper-parameters

3.2.1. Regulating learning rates. Learning rate which is a factor leading up to the gradient at the time of applying the SGD algorithm is of great relevance. Picture stepping down a U-shaped valley from a side of the valley to the nadir of it. In case the steps are extraordinarily massive, it would walk from one side directly to another and then return, so on and so forth. If it is so small, it will probably drop down into some tiny patches as it goes, because the path will invariably be uneven (local optimum), and having fallen down into those patches, it will never make it out of that pit if the steps stay the same width. Back to the model in this paper, here are author’s notes when using it, fixing the other parameters, and adjusting the learning rate, Table 4 is the experimental records with different learning-rate.

Table 4. Experimental records with different learning-rate.

n-kerns	Batch size	Pool-size	Learning-rate	Test error (%)	Validation error (%)
[20,50]	56	(2,2)	0.1	82.4	97.5
[20,50]	56	(2,2)	0.01	5	15
[20,50]	56	(2,2)	0.05	2.5	5

DOI: 10.54254/2755-2721/21/20231137

Note that there are only 56 images in both the validation and test sets, which means that only one or two images were identified incorrectly, which is still good. Finally, the author set the learning rate to 0.05.

3.2.2. Adjusting batch size. Because author use the minibatch SGD algorithm for optimization, the data is input batch by batch to the CNN model and then the calculation of the average loss for this batch of samples, i.e., cost function is the mean of the total number of examples. Batch size represents the quantity of examples comprised within one batch, and apparently, batch size has an effect on how well and how fast the model can be optimized. Returning to the model in this paper, firstly, since the training dataset is 392 and both the valid dataset and test dataset are 49, it would be better to have the batch size as a factor of 49, otherwise, some samples would be wasted. Here are the notes from author's experiments, fixing the other parameters and changing batch size. The time that set batch size equals to 1, 2, 5, 10, and 20, meanwhile validation error was still 97.5 percent but did not drop. The author find that the sample type covering is not big enough, for breakdown by classes, meanwhile 10 samples of each class are consecutively listed at the time that batch size amounts to 20, in fact it consists of just two classes, therefore optimization will be extremely difficult. Therefore, finally, the authors make the batch size equals to 56, and this adjust to the valid and test dataset size, due to the fact that the initial dataset is not big enough. Generally, the authors would not make the batch size equals to the valid and experimental dataset size.

3.2.3. n-epochs. N-epochs are the maximum number of training steps, e.g. If it is set to 200, then the training process will traverse the dataset at most 200 times and the program will stop when it has traversed the dataset 200 times. n epochs are equivalent to a control parameter that stops the program and does not affect the degree of optimization or the speed of the CNN model. But is simply a parameter that controls the end of the program.

3.2.4. n-kerns. Theoretically, the quantity the convolution kernel actually stands for means the quantity of traits, and the more traits which can be retrieved, the more correct the ultimate categorization possibilities will be. Nevertheless, there are still a lot of traits (plenty convolutional kernels) that add size to the parameters and computational sophistication, and sometimes more convolutional kernels are not better but should be determined according to the specific application object. Therefore, the author thinks that although CNN claims to automatically extract features, eliminating the need for complex feature engineering, many parameters such as the n-kerns here still need to be adjusted, and some" manual" work is still required. Here are author's experimental records, fixed batch size=56, learning rate=0.05, pool-size= (2, 2), in Table 5.

Table 5. Experimental records with different parameters.

n-kerns	Validation error (%)	Test error (%)	epoch
[20,50]	2.5	5	36
[10,30]	5	5	46
[5,10]	5	7.5	38

3.2.5. Pool-size. Pool-size is set to (2,2) in this program, i.e., max pooling 1 pixel from a 2*2 area, i.e., keeping 4 and a pixel to 1 pixel. The size of the face image in this example is 48*48, so (2,2) is reasonable for such a small image.

4. Conclusion

This study aims to address the problem that the accuracy of convolutional neural networks in expression recognition decreases with additional network deepness. The authors propose a combination of ResNet and SVM algorithms. Unlike other deep learning algorithms that rely on facial expression features in expression recognition, the algorithm in this paper uses a deep network to extract higher-order features directly. In addition, the model increases the network depth, and meanwhile decreasing parameters of

DOI: 10.54254/2755-2721/21/20231137

the network. The efficiency is maintained while improving the performance of the model. Moreover, the introduction of SVM effectively improves the recognition efficiency. The strength of SVM lies in solving small sample, non-linear and high-dimensional regression and binary classification problems. The results of the study indicated that recognition rate of the algorithm presented in this article is superior to traditional expression recognition algorithms. This in turn verifies viability and validity aspects concerning the algorithm. The model has good generalization ability in expression recognition. To further improve the model performance, the authors will focus on the impact of different SVM classifiers on the experiments in the future.

References

[1] UNCTAD 2019 The impact of rapid technological change on sustainable development
[2] Lindeberg T 2012 Scale invariant feature transform p 10491
[3] LeCun Y Bottou L Bengio Y et al 1998 Grandient-based learning applied to document recognition IEEE Pres 86(11): pp 2278-2324
[4] Krizhevsky A Sutskever I Hinton G 2017 ImageNet classification with deep convolutional neural networks Communications of the ACM 60(6): pp 84-90
[5] Taigman Y Yang M Ranzato M et al 2014 DeepFace: closing the gap to human-level performance in face verificatio IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE pp 1701-1708
[6] Sun Y Wang X Tang X 2015 Deeply learned face representations are sparse, selective, and robus IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE pp 2892-2900
[7] He K Zhang X Ren et al 2016 Deep residua for image recognitio IEEE Conference on Compu and Pattern Recognition (CVPR) IEEE pp 770-778
[8] Notley S Magdon-Ismail M 2018 Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and k-nearest neighbor classifier arXiv preprint arXiv:1805.02294
[9] Saifaldeen H Thanoon K 2021 A comprehensive study on high performance malware classifiers based on machine learning algorithm Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12(14): pp 4928-4938
[10] Kim S Yu Z Kil R et al 2015 Deep learning of support vector machines with class probability output network Neural Networks 64: pp 19-28
[11] Shi Y Cheng K Liu Z 2018 Segmentation of hippocampal subfields by using deep learning and support vector machin Journal of Image and Graphic 23(4): pp 0542-0551
[12] Zhang K Zhang Z Li Z et al 2016 Joint face detection and alignment using multitask cascaded convolutional network IEEE Pres 23(10): pp 1499-1503