Dynamic Facial Expression Recognition Based on Trained Convolutional Neural Networks

Ming Li[1,2,3] and Zengfu Wang[1,2,3](✉)

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, China
[2] University of Science and Technology of China, Hefei, Anhui, China
lm910415@mail.ustc.edu.cn, zfwang@ustc.edu.cn
[3] National Engineering Laboratory for Speech and Language Information Processing, Hefei, China

Abstract. Recently, dynamic facial expression recognition in videos receives more and more attention. In this paper, we propose a method based on trained convolutional neural networks for dynamic facial expression recognition. In our system, we improve Deep Dense Face Detector (DDFD) developed by Yahoo to reduce training parameters. The LBP feature maps of facial expression images are selected as the inputs of the designed network architecture which is fine-tuned on FER2013 dataset. The trained network model is considered as a feature extractor to extract the features of inputs. In an image sequence, the mean, variance, maximun and minimum of feature vectors over all frames are calculated according to its dimensions and combined into a vector as the feature. Finally, Support Vector Machine is used for classification. Our method achieves a recognition accuracy of 53.27% on the AFEW 6.0 validation set, surpassing the baseline of 38.81% with a significant gain of 14.46%. The experimental results verify the effectiveness of our method.

Keywords: Dynamic facial expression recognition · Face detection
Convolutional neural networks · Local Binary Patterns
Support Vector Machine

1 Introduction

With the rapid development of the biometric identification technology, facial expression recognition has attracted much attention. Automatic facial expression recognition has been an active research topic in computer vision and pattern recognition. It has enormous potential for development and can be used in intelligent human-computer interaction, mass entertainment, safe driving, medical assistance, online education, etc. Facial expression recognition aims to classify a given facial image or video into six prototypical categories (angry, disgust, fear, happy, sad and surprise). Automatic, accurate and real-time dynamic facial

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 218–226, 2018.
https://doi.org/10.1007/978-3-030-03335-4_19

expression recognition is still a challenge due to the complexity and variation of facial expressions.

Traditional facial expression recognition methods are mostly based on hand-crafted features. It takes a lot of time to design features. There are three main streams in the current research on facial expression recognition: geometry-based, texture-based and hybrid-based. The appearance of deep learning introduces facial expression recognition to a new stage. Specifically, convolutional neural networks have made breakthroughs in object detection, image recognition and many other computer vision tasks. It has become the prevalent entry solutions in recent facial expression recognition.

However, most of facial expression recognition methods have been performed on laboratory controlled data, which pooly represents the environment and conditions faced in real-world situations. To promote the development of emotion recognition under uncontrolled conditions, the Emotion Recognition in the Wild (EmotiW) challenge [3–7] has been held from 2013 by ACM. The goal of EmotiW challenge is to provide a common platform for evaluation of emotion recognition methods in real-world conditions. In the past five years, a large number of methods have been proposed for dynamic facial expression recognition. Yao [19] analyzed spontaneously expressed emotions from the perspective of making a deep exploration of expression-specific AU-aware features and their latent relations. [9] proposed a hybrid network that combines recurrent neural network and 3D convolutional neural networks in a late-fusion fashion. HoloNet combined CReLU, residual structure and inception-residual structure to produce a deep yet computational efficient convolutional neural network for emotion recognition in the wild [18]. Hu presented a new learning method named Supervised Scoring Ensemble (SSE) [12] for advancing EmotiW challenge with deep convolutional neural networks.

In this paper, we propose a novel dynamic facial expression recognition method based on trained convolutional neural networks. The primary contributions of this work can be summarized as follows: 1. Based on the multi-view face detection algorithm developed by Yahoo [10], we improve the performance and reduce the training parameters by changing its network architecture and establishing a more appropriate face database. 2. The Local Binary Patterns (LBP) [15] feature maps are selected as the inputs of the designed network architecture. We try to explore the performance of network using different types of input. 3. The trained convolutional neural network model is considered as a feature extractor to capture the features of the inputs. Then, in an image sequence, the mean, variance, maximum and minimum of feature vectors over all frames are calculated according to its dimensions and combined into a vector as the feature. Finally, Support Vector Machine (SVM) [2] is utilized for classification. Our method only uses a single network model and achieves a recognition accuracy of 53.27% on the AFEW 6.0 validation set, surpassing the baseline of 38.81% with a significant gain of 14.46%.

The rest of this paper is organized as follows. Section 2 describes the proposed method, including preprocessing, network architecture, fine-tune training,

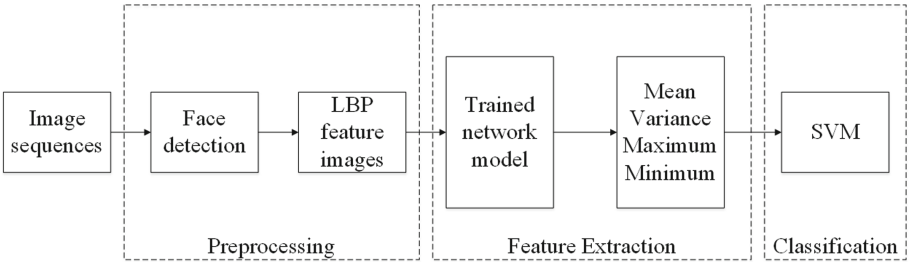

Fig. 1. Framework of the proposed dynamic facial expression recognition method.

feature extraction and facial expression classification. In Sect. 3, we evaluate our scheme and compare it with other methods. Finally, conclusions are presented in Sect. 4.

2 The Proposed Method

The framework of the proposed method is shown in Fig. 1. It consists of three parts: preprocessing, feature extraction and classification. In what follows, we will give the detailed descriptions for each part.

2.1 Preprocessing

The preprocessing procedure consists of three steps. First, we use a improved DDFD multi-view face detector to locate target faces from all frames in a facial expression sequence. And the frames without faces are directly discarded. Second, the detected facial images are resized to match the input size of our designed network architecture. Third, the facial images are converted to its corresponding LBP feature maps, which are feed to the network.

Traditional face detection algorithms only detect frontal or close-to-frontal faces. However, humans have various head posture changes in the wild, which requires multi-view face detector. So we adopt DDFD face detection algorithm and make some improvements. The original DDFD algorithm only uses a single model based on deep convolutional neural networks. And it has minimal complexity and does not require additional components for segmentation, bounding-box regression, or SVM classifiers.

The network architecture of DDFD algorithm is similar to AlexNet, which contains relatively large convolution kernels, such as 5×5, 11×11. They bring more weight parameters. Therefore, we change the network architecture to reduce the amount of parameters. Specifically, we replace a 5×5 kernel with two concatenated 3×3 kernels, and replace a 11×11 kernel with five concatenated 3×3 kernels (see Fig. 2). At the same time, 1×1 kernel can decrease the feature maps and achieves cross-channel information integration. This design can reduce the parameters and improve the nonlinearity of the network. Compared

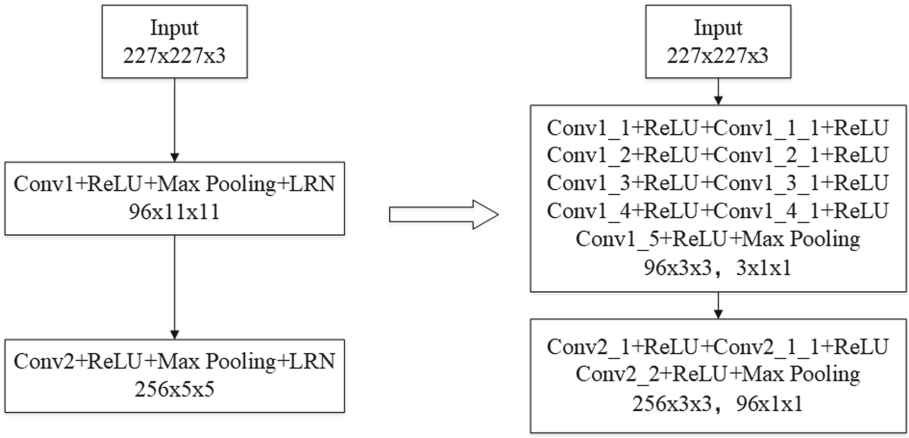

Fig. 2. The improved network architecture in relatively large convolution kernels.

with original network, the parameters of the first and second convolutional layer are reduced by 59% and 24%, respectively. Moreover, it increases the depth of the network and contributes to learn features.

In addition, the number of positive and negative samples used by DDFD algorithm is quite different. And the positive samples are randomly sampled sub-windows of the images if they have more than a 50% IOU (intersection over union) with the ground truth, which is insufficient to express faces. So we build a new face detection dataset. It contains 91067 face images and 96555 non-face images. The face images are cropped by IOU>= 0.65 from AFLW [13] dataset. And the non-face images are sampled and selected from the PASCAL VOC2012 dataset.

2.2 Network Architecture

In our work, we only train a single network architecture: VGG-Face [16]. It is a network model based on convolutional neural network proposed by the Institute of Visual Studies at Oxford University for face recognition. Figure 3 gives the details of the network architecture that contains 13 convolutional layers and 3 fully connected layers. Each convolutional layer is followed by ReLU activation function and a max-pooling layer, which adds nonlinearity to the network and reduces the dimension of feature maps. And all the convolutional kernels are 3×3. Besides, we use dropout in the first two fully connected layers to reduce over-fitting of the network. The input to the network is a LBP feature map of size 224×224 with the average face image (computed from the training set) subtracted.

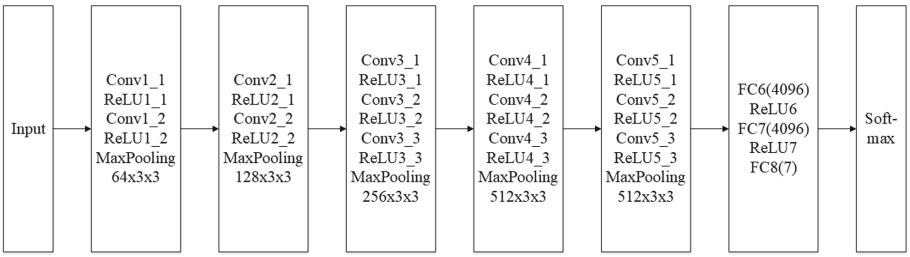

Fig. 3. The network architecture.

2.3 Fine-Tune Training

Due to the relatively small number of samples in AFEW 6.0 dataset [8], the network is fine-tuned on FER2013 dataset [11]. The trained VGG-Face model is regarded as pre-training model, whose weight parameters are used to initialize the network. And the initial learning rate of the network should not be set too large. The learning rate is initially set to 10^{-4} and decreased by factor of 10. Optimisation is by stochastic gradient descent using mini-batches of 64 samples and momentum coefficient of 0.9. The coefficient of weight decay is set to 5×10^{-4}, and dropout is applied with a rate of 0.5.

2.4 Feature Extraction

Different from traditional feature extraction methods, the features are extracted based on trained convolutional neural network (see Fig. 4). For an image sequence, the LBP feature map of each frame is feed to the trained network model for extracting features. Then, we choose the feature maps of fully connected layer 6 (fc6) from the network as the feature vector of the input image. The dimension of the feature vector is 4096.

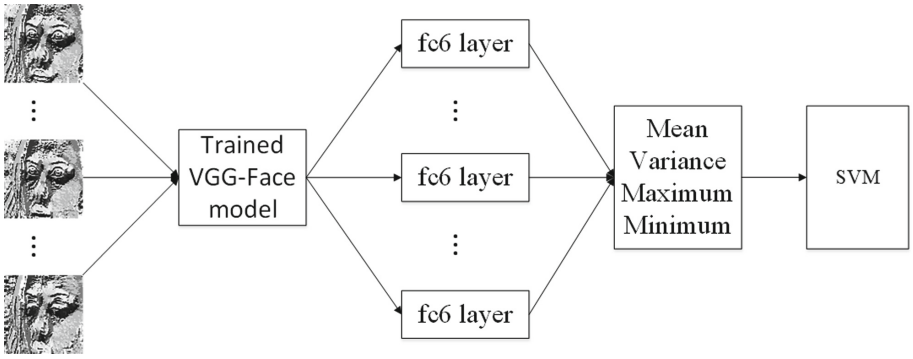

Fig. 4. The process of feature extraction.

After the above process, the set of feature vectors from all frames can be obtained. Then, we attempt to make use of these frame-level feature vectors to represent the image sequences. In an image sequence, the mean, variance, maximum and minimum of feature vectors over all frames are calculated according to its dimensions and combined into a vector. Its dimension is 16384(4096×4). Finally, the vectors are normalized as the feature representation.

2.5 Facial Expression Classification

After extracting features, we train a Support Vector Machine (SVM) to give each sequence one of the seven emotion classes. Facial expression recognition is a multi-class problem, so we use LIBSVM with the Radial Basis Function (RBF) kernel for classification. To evaluate the performance of our method, we use 5-fold cross-validation, 10-fold cross-validation, 15-fold cross-validation and leave-one-out cross-validation schemes in our work.

3 Experiments

3.1 Datasets

In order to train the network model and evaluate the performance of the proposed method, we adopt two datasets: FER2013 dataset and AFEW 6.0 dataset.

FER2013 dataset. It's a large and publicly available facial expression dataset. The dataset contains 28709 training images, 3589 validation images and 3589 test images. All images are grayscale and have a resolution of 48×48 pixels. Each image is labeled with one of seven expressions: anger, disgust, fear, happy, sad, surprise or neutral. In our experiments, the dataset is used to fine-tune the network model.

AFEW 6.0 dataset. The Acted Facial Expressions in the Wild (AFEW) 6.0 Dataset is the official dataset provided by the EmotiW 2016. It consists of 1749 video clips, which are split into three parts: 773 samples for training, 383 samples for validation and 593 samples for test. All video clips are collected from Hollywood real movie records and reality TV clips. Therefore, there are numerous variations in head pose movements, lighting, background, occlusion, etc. Only samples for training and validation have emotional labels, which are categorized into seven classes: anger, disgust, fear, happy, sad, surprise and neutral. Due to the lack of emotional labels in test set, we conduct our experiments on the training and validation sets.

3.2 Experimental Results

The experiments consist of two parts: fine-tuning training experiment and dynamic expression recognition experiment. Then, we analyze the experimental results of the two parts.

Fine-Tuning Training Experiment. Table 1 shows the recognition accuracy fine-tuned on FER2013 dataset. Compared with other methods, our experimental result is also acceptable. We achieve a recognition accuracy of 71.28% on FER2013 test set. The humans accuracy on this dataset is around 65.5%, and our method exceeds 5.78% of it.

Table 1. Recognition accuracy fine-tuned on FER2013 dataset

Method	Recognition accuracy (%)
Tang [17]	71.20
Yu [20]	72.03
Mollahosseini [14]	66.4 ± 0.6
Ours	71.28

Dynamic Facial Expression Experiment. We combine the training set and validation set of AFEW 6.0 dataset. Then, we use 5-fold cross-validation (5-fold-CV), 10-fold cross-validation (10-fold-CV), 15-fold cross-validation (15-fold-CV) and leave-one-out cross-validation (LOO-CV) schemes to conduct experiments, respectively. Table 2 presents the recognition accuracy using different schemes. As shown in Table 2, when 15-fold-CV is adopted, the recognition accuracy is highest.

Table 2. The recognition accuracy using different schemes.

Schemes	Recognition accuracy (%)
5-fold-CV	52.17
10-fold-CV	52.83
15-fold-CV	53.75
LOO-CV	52.76

In order to compare with other methods, we use SVM to train facial expression classifier on AFEW 6.0 training set. Then, the classifier is utilized to test the samples from AFEW 6.0 validation set. We get a recognition accuracy of 53.27% which surpasses the baseline of 38.81% with a significant gain of 14.46%. The experimental results compared with other method are shown in Table 3. The recognition accuracy of our method is better than the first and second place in EmotiW 2016 challenge, but we still have a certain gap from [1]. The reason may be that they trained a better network model using additional dataset. In addition, we use original expressional images as the inputs of the network, and obtain a recognition accuracy of 51.57%. It means that the LBP feature maps are helpful for facial expression recognition, which also explains the impact of different inputs on network performance.

Table 3. Comparisons with different approaches on AFEW 6.0 validation set.

Method	Recognition accuracy (%)
Baseline	38.81
Fan [9]	51.96
Yao [18]	51.96
Bargal [1]	59.42
Ours	53.27

4 Conclusions

In this paper, we propose a novel method for dynamic facial expression recognition in the wild. The original DDFD face detection algorithm is improved in network architecture to reduce the parameters and increase the nonlinearity of the network. We use the LBP feature maps of expressional images as the inputs of the network architecture which is fine-tuned on FER2013 dataset. Then, the trained network model is used to extract feature of one sequence. The experimental results verify the effectiveness of our method. In the feature, we will select other network models to extract different features to enhance the representation of the facial expressions. Besides, we will integrate audio information, language information and behavioral information to help facial expression recognition.

Acknowledgement. This work is supported by National Natural Science Foundation of China (No: 61472393).

References

1. Bargal, S.A., Barsoum, E., Ferrer, C.C., Zhang, C.: Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 433–436. ACM (2016)
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
3. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: EmotiW 5.0. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 524–528. ACM (2017)
4. Dhall, A., Goecke, R., Joshi, J., Hoey, J., Gedeon, T.: EmotiW 2016: video and group-level emotion recognition challenges. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 427–432. ACM (2016)
5. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 461–466. ACM (2014)
6. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 509–516. ACM (2013)

7. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426. ACM (2015)
8. Dhall, A., et al.: Collecting large, richly annotated facial-expression databases from movies (2012)
9. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450. ACM (2016)
10. Farfade, S.S., Saberian, M.J., Li, L.J.: Multi-view face detection using deep convolutional neural networks. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 643–650. ACM (2015)
11. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013. LNCS, vol. 8228, pp. 117–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42051-1_16
12. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 553–560. ACM (2017)
13. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE (2011)
14. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
15. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
16. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. vol. 1, p. 6 (2015)
17. Tang, Y.: Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239 (2013)
18. Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: HoloNet: towards robust emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 472–478. ACM (2016)
19. Yao, A., Shao, J., Ma, N., Chen, Y.: Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 451–458. ACM (2015)
20. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442. ACM (2015)