Micro-expression Recognition Using Enriched Two Stream 3D Convolutional Network

He Yan*
College of Artificial Intelligence
Liangjiang Chongqing University of Technology
Chongqing, P.R. China
yanhe@cqut.edu.cn

Lei Li
College of Artificial Intelligence
Liangjiang Chongqing University of Technology
Chongqing, P.R. China
1034273169@qq.com

ABSTRACT

Face micro-expression recognition in video is an active field in computer vision research and its difficulty lies in transience in mic-expression motion and limited mic-expression databases. Particularly, the unevenness of the data set samples caused by the difficulty of capturing some types of micro-expressions will result categories with a small sample size contain too few features and it will be hard to improve this kind of category recognition rate. This can be proved by the low accuracy shown by the most advanced methods. In this paper, for spatial and temporal feature enrichment, we amplify the action amplitude of micro expressions by phase - based video amplification and use optical flows to describe the action characteristics of micro expressions. We propose a two-stream 3D convolutional network to extract both amplified optical flows and original frames. The experiment was carried out on CAS (ME)^2 micro-expression dataset. The proposed model is superior to existing methods especially in small sample categories.

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Computer vision • Computer vision problems • Object recognition

KEYWORDS

Micro-expression recognition, Two-stream 3D convolution network, Phase-based video amplification, Optical flow

1 Introduction

Micro–expressions are temporary and involuntary facial expressions that occur when real emotions are hidden in the heart of a person. Understanding these micro–expressions can help us recognize deception and understand a person's real psychological state. In recent years, some works using computer vision and machine learning algorithms to realize automatic micro-expression recognition have been proposed.

Before 2015, researchers used traditional hand-designed methods for feature extraction of micro-expressions recognition, Pfister et al. [1] were the first to study the recognition of micro-expressions who put forward the LBP-TOP method of extracting micro-expressions features, then new operators based on this constantly keep coming in to solve this problem [2-4]. However, manual design features extraction is difficult to improve its recognition performance due to its limited feature expression ability.

In recent years, people have tried to use the neural network method for micro expression recognition and achieved higher recognition. The quality of the dataset is very import for the feature extraction of neural network. However, due to the low motion range and short duration of each micro-expression and many researches have not deliberately considered the categories of small samples question for scarcity of data in the micro-expression dataset. Micro-expression recognition poses challenges to the usage of deep neural network. In order to overcome the limitations of the existing techniques, this paper proposes an enriched two-stream 3D convolution network which use 3D convolution neural network to take both original frames of videos and amplified optical flows as input to extract features. This paper is structured as follows: Part2 presents a detailed review of related work on micro-expression recognition. Part3 proposes the proposed enriched two-stream 3D convolutional network. Part4 summarizes the experimental settings including data sets used and hyperparameters and a detailed discussion on the experimental results and analysis are followed in Part5. Finally, Part6 concludes the paper with remarks.

2 Related Works

In this section, we briefly review and discuss the latest methods for micro-expressions recognition.

2.1 Manual Extraction of Features

Hand-designed features based approaches for micro-expression recognition were started early. Since research of Pfister, many researchers have proposed new methods of manual extraction. In order to extract micro-expression motion feature, Literature [5]

CSAE 2020 October 20-22, 2020, Sanya, China

ACM ISBN 978-1-4503-7772-0/20/10 $15.00
https://doi.org/10.1145/3424978.3425116

proposed optical strain weighted features and optical strain amplitude feature extraction methods, which use optical strain to characterize motion information, thereby calculating tiny facial muscle movements, and using temporal and spatial information to recognize micro-expressions. Based on optical flow method, Literature [6] used a simple and effective Main Direction Average Flow Characteristics (MDMO). Literature [7] used a fuzzy histogram of optical flow orientation (FHOFO) to extract micro-expression features. They proved that optical flow can better characterize the motion characteristics of micro-expressions than original image. Literature [8] used euler video amplification technology to amplify micro-expression motion, and then used LBP-TOP method to extract the amplified features. It showed good performance in CASME Ⅱ dataset. It was proved that EVM method can improve the performance of micro-expression recognition. Literature [9] analyzed the information between the two frames by calculating the direction, the optical flow amplitude, and the optical flow strain between the initial frame and the maximum frame. However, all these manual feature extraction methods cannot break through the limitation that they can only extract superficial features of micro-expression.

2.2 Learning-based Methods

With the improvement of computing power, deep learning methods far surpass traditional methods in image processing and video analysis. Researchers are also trying to use deep learning methods for micro-expression recognition. Literature [10] first introduced the deep learning method into the recognition of micro expressions and effectively extracted the features of micro expressions through feature selection.

However, the insufficient data in the datasets of micro-expression recognition is a common problem, which is also the reason for the unsatisfactory accuracy of micro-expression recognition. A major problem with any technology based on deep learning is that it requires enough samples. However open micro-expression datasets currently includes USF-HD, SMIC, CASME don not have sufficient quantity and they all have categories with small samples. It is challenge to learn in small samples. Literature [11] proposed a method to extract optical flow characteristics from micro-expression video frames to classify micro-expression. The extracted optical flow features are fed into the CNN model for expression classification. Literature [12] designed a two-stream shallow convolutional neural network and input the optical flow feature map to neural network. But robust and comprehensive features cannot be obtained just using a single optical flow feature map. Literature [13] used the network architecture of CNN+LSTM. CNN was used to encode the spatial features of facial expressions in different expression states. LSTM (Long Short-term Memory) learned the temporal state change information of the micro-expression spatial features. Literature [14] proposed an Enriched Long-term Recurrent Convolutional Network (ELRCN) that first encoded each micro-expression frame into a feature vector through CNN modules, then predicts the micro-expression by passing a LSTM module. They used the optical flow as the characteristic enrichment of the input. But LSTM cannot describe the spatial characteristics of micro-expression very well. In order to simultaneously extract the temporal and spatial features of micro-expression. Literature [15] proposed to use 3D convolutional neural network to recognize micro-expressions. They used consecutive video frames as input to the neural network. But due to the weakness of the micro expression movement, continuous frame sequences cannot commendably represent the movement of micro-expressions.

3 Proposed Method

To break through the limitations of current technology, we need carry out characteristic enrichment coding for subtle facial changes in micro-expression datasets. This paper proposes an enriched two stream 3D convolutional network which uses 3D convolutional layers to extract temporal and spatial features of micro-expression and enriches the spatial features by superimposing the input channels. And the phase - based video magnification algorithm is used as a pre-process method.

The main contributions of this paper are summarized as:

(1). In order to increase the characteristics of the micro-expression data set, the selected micro-expression data set is pre-processed, the micro-expressions in the dataset are first zoomed in, and then the optical flows between frames are extracted.
(2). In order to reduce the noise of the amplification algorithm,we use the phase-based amplification algorithm.
(3). To extract both optical flow and original image features, this paper adopts a two-stream 3D convolutional neural network.

Figure 1 shows the complete structure of the method in this paper:

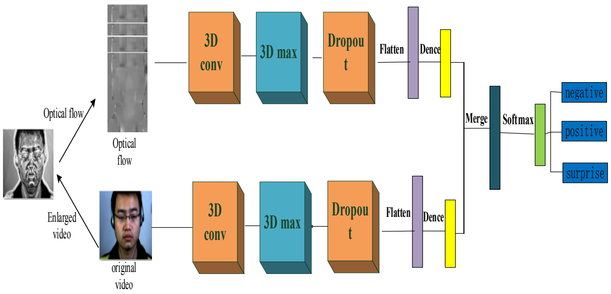

Figure 1: The Proposed Network and Pre-processing Method.

3.1 Pre-processing

3.1.1 PBVM. Firstly, the phase-based video magnification algorithm is used to magnify micro-expression video. Researchers have used euler video magnification algorithm to magnify micro-expressions. Experiments show that the amplification algorithm is effective to improve the recognition rate. However, euler video amplification technique is a linear micro motion amplification method. Although the algorithm is simple and fast, it also has obvious defects. Its magnification is limited, and it increases the amplitude of the small motion signal but also causes the linear increase of the noise. So this paper use phase-based video amplification technology (PBVM) [16]. The improved method uses complex manipulable pyramid to filter video sequences in the space frequency domain and enhances the motion of the pixels by changing the phase of the sinusoidal pattern that makes up the image. Because there are a lot of microscopic facial movement in the videos, but we're interested in those local actions. In order to magnify these local movements, we adopt a complex directional pyramid, which uses the local fourier transform to decompose each frame image into spatial structure images with different scales and directions. In this method, the synchronization amplification of noise is not caused when the motion is amplified, and a larger amplification factor is also supported.

3.1.2 Optical-flow. We extracted the magnified video for optical flow, which was proposed by Gibson in 1950. Optical flow refers to the velocity of mode motion in the image, which is a two-dimensional instantaneous velocity field. The optical flow field corresponds to the sports field under ideal conditions, and it is found that the optical flow field is more robust in feature extraction than the ordinary sports field. In 1981, Horn and Schunck [17] connected the two-dimensional velocity field with the gray scale and introduced the optical flow constraint equation. In this paper, enlarged image optical flow constraint equation can be described:

$$\hat{i}_x u + \hat{i}_y v + \hat{it} = 0 \tag{1}$$

where, $\hat{i}_x$, $\hat{i}_{\mathrm{y}}$, $\hat{it}$ represents the partial derivatives along x , y and time direction of the grayscale of the pixel in the enlarged image respectively, and (u, v) is the optical flow vector to be obtained. The variables are two. There is only one constraint equation, and another constraint: equation need to be introduced. Literature [18] shows that the L^1 norm in the variational energy functional can be introduced into the optical flow constraint to better maintain the edge information of the image on the optical flow image. Therefore, this paper uses the $TV-L^1$ optical flow method to perform optical flow extraction on the enlarged video. In order to form three-dimensional flow images, horizontal and vertical flow images are connected in series.

3.2 Proposed Learning Method

Since the micro-expression dataset is composed of short videos, this paper use 3D convolutional neural network to recognize the pre-processed micro-expression optical flow and consecutive frames. Compared with the traditional 2D neural network, the 3D convolutional neural network is a deep learning network that can extract time-domain features. 3D convolution can make full use of the data space and the information between channels, and can capture the information between multiple frames. To enrich the characteristics of small samples, this paper uses a two-stream 3D convolutional neural network to extract the continuous frame and optical flow information simultaneously. It consists of 3D convolution layer, 3D pooling layer, full connection layer, activation function and dropouts. 3D convolutional layer extracts spatiotemporal features; 3D pooling layer gradually reduces the size output of 3D convolutional layer while retaining important features; the use of dropout reduces the overfitting of the model to the training sample. After the layer of flatten, it compresses the features down to 128 dimensions and merges two networks. The soft-max layer is used to generate the class scores for the classes of the dataset being used. The input has three dimensions w*h*c which means width, height, and channels of the input. Considering that the number of data in experimental data set is small, we do not choose many network layers. We selected 8 consecutive frames of micro-expression video which includes the start to end frames of micro expressions. So the extracted optical flows become 16 consecutive frames. Table 1 shows the detailed information of each layer of the neural network:

Table 1: Proposed Network Model Structure.

Layer Type	Filter Size	Output Dimension
conv3d_1	3*3*15	32*64*64*8
max_pooling3d_1	3*3*3	10*21*21*8
dropout_1		10*21*21*8
flatten_1		35280
dense_1		128
conv3d_2	3*3*15	32*64*64*16
max_pooling3d_2	3*3*3	10*21*21*16
dropout_2		10*21*21*16
flatten_2		70560
dense_2		128
merge		256
soft-max		3

4 Experimental Setting

This section discusses the follow-up experimental settings of this work.

Since there are many micro-expression videos that are not suitable for 3D convolution kernel. So we choose CAS (ME)^2 [19] which is the latest version of the CASME series of datasets on facial micro-expressions containing 206 videos. This dataset contains 3 classes (negative, positive, surprise). Note that we have only used micro-expression videos that have more than 8 frames to maintain the consistency over the data which includes 164 short videos. We use 80% of the data set as the training set and 20% as the validation set. Table 2 shows the details of the experimental dataset.

This paper first use PBVM to amplify the micro-expressions, and the selected magnification factor α =20, and then use open-cv to extract 8 consecutive frames and convert them into 8 frames of optical flow images along the x and y direction. To minimize errors, the image is uniformly cut into 64*64 dimensions and normalized. We implemented our model on Keras using Tensorflow as the back end. The model was trained and tested on a GPU server using an NVIDIA Quadro RTX 8000 graphics processor. The classification cross entropy loss function is used and the default learning rate scheduling is adopted. In addition, networks trained 100 epochs with a batch size of 8. The input dimensions of the network model are 64*64*8 and 64*64*16 respectively.

Table 2: The Number of Samples for Each Category in the Experimental Data Set.

negative	67
Positive	69
Surprise	28
Total	164

5 Experimental Results and Discussions

This section presents the experimental results. In order to analyze the overall recognition effect and the effect of small sample categories of our method, this paper makes a rigorous comparison with the most advanced deep learning methods from recent papers [20, 21]. We analyze accuracy standard deviation of the proposed method to analyze the advantages of the model in the small sample micro-expression dataset. In order to effectively evaluate the network, we use other methods to conduct a comparative test of the data set. We use f1 scores and recall rates as evaluation indicators and obtain the recognition rate of each category through the confusion matrix. The experimental results are showed in Table 3:

Table 3: Comparison of Our Methods and Other Convolutional Neural Network.

Method	Accuracy of validation set	f1_score	Recall rate
CNN_RNN	0.58	0.57	0.55
One stream 3D CNN	0.57	0.58	0.56
Our method	0.66	0.63	0.60

Obviously, it can be seen from the experimental results that our method is overall superior to other works. The first two methods just have one-stream network and just take the original frame sequences as the input of the network. It cannot extract deep features of micro expressions. In our method, we use a two-stream 3D convolutional neural network with optical flow and original frames as input which can extract more features than the single-stream network. Due to optical flow can better describe micro expression movement and has better robustness, our method can obtain deeper motion information of the micro-expression. Since the surprise class is the least numerous class in the dataset which only has 28 videos, and surprised categories are easily confused with other categories like positive. This kind of category is difficult to recognize. you can see from the confusion matrix in Figure 2. The first two methods have less than a 30% recognition rate which means they can't totally recognize micro-expressions when the number of the category is not enough and it is easy for the network to confuse the surprise category with other categories. In our method, we use PBVM as micro-expression pre-processing method. It does help for expanding the difference of different categories by amplification of micro-expression movement. For further analysis, we used different preprocess methods to conduct a comparative experiment of neural network micro-expression recognition to analyze the impact of our data preprocess methods. We extracted the original micro-expression optical flows and the optical flows processed by euler video amplification algorithm to compare with our method. The experimental results are showed in Table 4:

Table 4: Comparison of Different Pre-processing Methods.

Input	Accuracy	F1_score	Recall rate
Raw video frame optical flow	0.58	0.58	0.56
EVM	0.60	0.61	0.56
PBVM	0.66	0.63	0.60

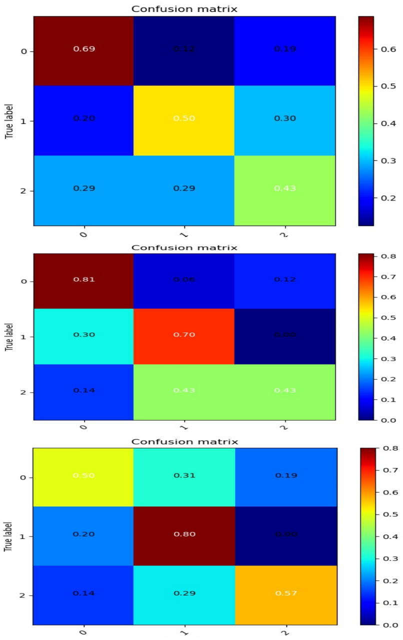

Figure 2: Confusion Matrix of Methods (CNN_RNN, One Stream 3D CNN, Our Method).

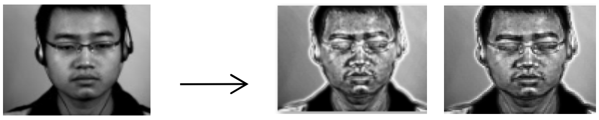

Figure 3: Video Amplification Algorithm Results (Original Frame, EVM, PBVM).

It can be seen from the experimental results that the proposed pre-processing method in this paper has the highest precision in the data set. First, the experiment proves that the amplification algorithm can effectively improve the recognition rate of micro expressions. It can be seen from Figure 3 that the EVM also amplifies the noise while amplifying the micro-expression action Because EVM is a just linear amplification algorithm which can only amplify the micro-expression video as a whole. It also amplifies noises like blinking motion. However, PBVM is a local amplification algorithm. We only extract the needed motion fragments to amplify. It doesn't amplify noise. It can be concluded that PBVM is a more ideal amplification algorithm for micro expression recognition task.

6 Conclusions

This paper proposes an enriched two stream 3D convolutional network to recognize micro expression in CAS (ME)^2. The micro-expression movements are amplified by PBVM, and optical flow is used as one of the inputs of the neural network. Experiments show that compared with the existing methods, the method proposed in this paper has a certain recognition rate improvement especially in categories with small sample.

Although this paper proposes an optimization strategy, the recognition performance is still limited due to the insufficient sample size of the data set. In order to further solve the problem of small samples, data enhancement methods should be considered in subsequent work.

ACKNOWLEDGMENTS

The National Key R & D Plan “Intelligent Robots” Key Project of P.R. China (Grant No. 2018YFB1308602), the National Natural Science Foundation of P.R. China (Grant No. 61173184), the Chongqing Natural Science Foundation of P. R. China (Grant No. cstc2018jcyjAX0694).

REFERENCES

[1] T Pfister, X Li, G Zhao and M Pietikainen (2011). Recognizing spontaneous facial micro-expressions. In 2011 International Conference on Computer Vision, pp. 1449-1456, IEEE.[2] S J Wang, H L Chen, W J Yan, Y H Chen and X Fu (2014). Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. Neural Processing Letters, vol. 39, no. 1, pp. 25-43.

[3] X Huang, S J Wang, G Zhao and M Piteikainen (2015) Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1-9.

[4] Y Guo, C Xue, Y Wang and M Yu (2015). Micro-expression recognition based on CBP-TOP feature with ELM. Optik, vol. 126, no. 23, pp. 4446-4451.

[5] Liong S T, See J, Phan R C W, et al. (2014). Subtle expression recognition using optical strain weightedfeatures. In 2014 Asian Conference on Computer Vision, vol. 47, pp. 160-172.

[6] Y Wang, J See, R C W Phan and Y H Oh (2015) Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. PloS one, vol. 10, no. 5, pp. e0124674.

[7] Happy S L and Routray A (2019). Fuzzy histogram of optical flow orientations for micro-expression recognition. In IEEE Transactions on Affective Computing.

[8] S Y Park, S H Lee and Y M Ro (2015). Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In Proceedings of the 23rd ACM International Conference on Multimedia, pp. 911-914.

[9] S T Liong, et al. (2016). Spontaneous subtle expression detection and recognition based on facial strain. Signal Processing: Image Communication, vol. 47, pp. 170-182,

[10] D Patel, X Hong and G Zhao (2018). Selective deep features for micro-apexnet on micro-expression recognition system. arXiv preprint arXiv:1805.08699.

[12] Khor H Q, See J, Liong S T, et al. (2019). Dual-stream shallow networks for facial micro-expression recognition. In 2019 IEEE International Conference on Image Processing (ICIP), pp. 36-40, IEEE.

[13] Kim D H, Baddar W J and Ro Y M (2016). Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In the 24th ACM International Conference on Multimedia, Feb. 1-4, Melbourne, VIC, Australia, pp. 382-386.

[14] H Q Khor, J See, R C W Phan and W Lin (2018). Enriched long-term recurrent convolutional network for facial micro-expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 667-674, IEEE.

[15] J Li, Y Wang, J See and W Liu (2019). Micro-expression recognition based on 3D flow convolutional neural network. Pattern Analysis and Applications, vol. 22, no. 4, pp. 1331-1339.

[16] A C Le Ngo, J See and R C W Phan (2016). Sparsity in dynamics of spontaneous subtle emotions: analysis and application. IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 396-411.

[17] B K Horn and B G Schunck (1981). Determining optical flow. in Techniques and Applications of Image Understanding. International Society for Optics and Photonics, vol. 281, pp. 319-331.

[18] T Pock, M Urschler, C Zach, R Beichel and H Bischof (2007). A duality based algorithm for TV-L 1-optical-flow image registration. In International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 511-518, Springer.

[19] F Qu, S J Wang, W J Yan, H Li, S Wu and X Fu (2017). CAS (ME) $^ 2$: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 424-436,

[20] S P T Reddy, S T Karri, S R Dubey and S Mukherjee (2019). Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks. In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, IEEE.

[21] H Q Khor, J See, R C W Phan and W Lin (2018). Enriched long-term recurrent convolutional network for facial micro-expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 667-674, IEEE.