Action recognition via Stacked Short Term Deep Features and Bidirectional Moving-Average Neural Network

*Jinsol Ha, Joonchol Shin, Hasil Park and Joonki Paik**
Graduate School of Advanced Imaging Science, Multimedia, and Film
Chung-Ang University
Seoul, Korea
jinsol5167@gmail.com, mbstel275@gmail.com, hahaha2470@gmail.com, paikj@cau.ac.kr*

Abstract—To recognize the actions, both spatio and temporal information are necessery. However, the temporal feature ex traction methods such as optical flow have high computational complexity. Hence, this paper presents, an action recognition method using a novel and simple spatial-temporal information. The proposed method consists of i) We generate RGB and differential images and extract the deep feature maps using image claasification network. ii) Generated deep features are concate nated by the proposed neural network and train with neural network. iii) Weights are given to important frames through a moving average. vi) Classify the probabilities by softmax function. The experimental result shows that the proposed short-term feature and bilateral moving average can improve the accuracy for an action recognition.

Index Terms—Action Recognition

I. Introduction

Intensive Action recognition research is essential to video analysis and automatic surveillance of intelligent surveillance systems. The proposed 2D convolutional neural net(2DCNN) can recognize the features of object effectively [1]. However, since only spatial information of the image is calculated, errors can occur when applied to action recognition. Ishan et al. proposed a raw spatio-temporal signals based approach to classify the videos [2]. However, its low performance makes it difficult to apply to action recognition that requires higher accuracy.

In this paper, we propose a novel RGB intensity variation based deep feature that can apply spatial-temporal information of input image. The proposed deep fearure and simple-neural net can recognize the action by combining spatial information and short-term pixel difference information between adjacent frames. Therefore, the proposed method can perform fast and accurate action recognition.

II. Proposed method

The proposed method extract the input image to 14 frames

$S_t =$

A. Generate the Differential Image

The proposed method to create differential image is follows:

$$I_t^{diff}(x, y) = I_{t+1}(x, y) - I_t(x, y), \quad (1)$$

where x and y are location of each pixel in the image. t represents the temporal parameter for each frame. Equation (1) uses the difference of pixel information values between frames so that the differential image includes motion information of the image.

B. Deep Features

Extracted feature maps calculated follow Fully-Connected layer:

$$V_t = F\,C_t(concat((F(I_t)), F(I_t^{diff}))), \quad (2)$$

where *concat* is the feature map concatenation operator, $F\,C_t$ represents t-th FC layer. F returns the last feature map of the convolutional layer of VGG16. The generated feature vector V_t is synthesized by the following the Bidirectional exponential moving average operation.

C. Bidirectional exponential moving average

In general, action is sequential. Therefore important infor mation is included in the middle of the image frame where the action is performed in the training video. By applying a bidirectional expoenetial moving average, the weight of which is lowered from the center frame to the anode, the weight of the image information of the center frame is increased. The bidirectional exponential moving average can be calculated recursively as follows:

$$V_t,\ t = a \quad \alpha \cdot V_t + (1 - \alpha) \cdot S_{t+1},\ t < a \quad \alpha \cdot V_t + (1 - \alpha) \cdot S_{t-1},\ t > a \quad , \quad (3)$$

for action recognition. In order to extract the spatial informa tion of the RGB images and temporal information of given frames, the composite product layer of VGG16 pre-trained with Image-Net is applied.

S_t, which represents the value of the Bidirectional exponen tial moving average, is divided symmetrically about the frame *a*. α is a value between 0 and 1. In this paper, the experiment was conducted with α as 0.9.

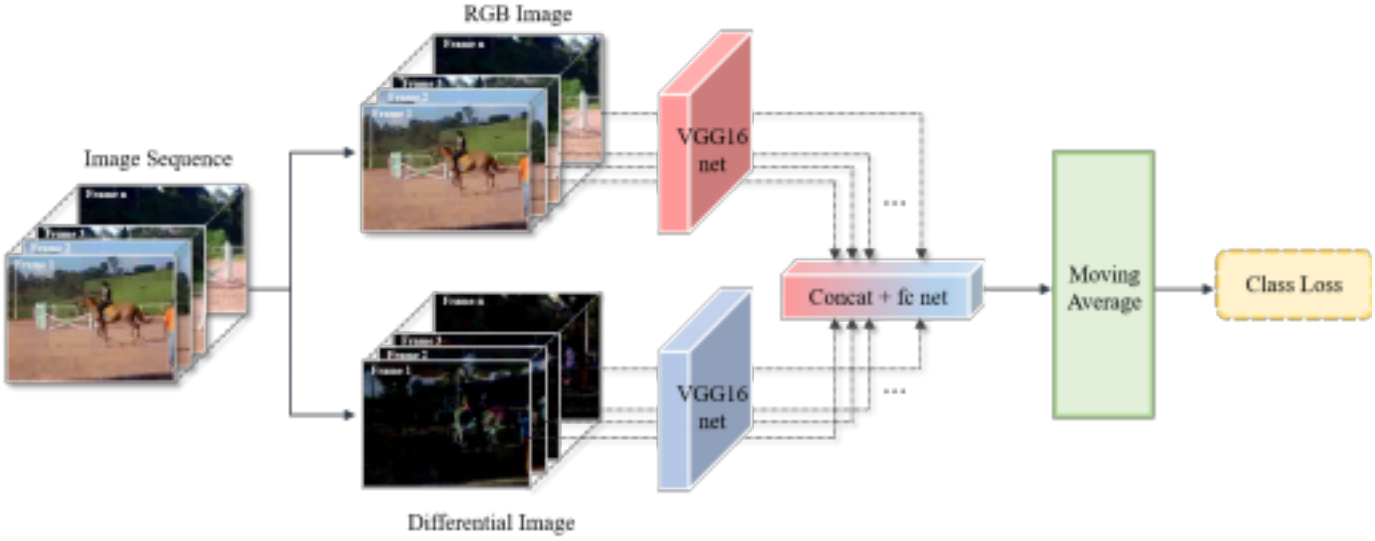

Fig. 1: The architecture of proposed method.

D. Loss function

Finally, to verify the performence of the proposed feature map, we desined simple neural network except fc layer. the loss function of the neural layer is as follows:

$$c = \text{sum}[(label - \hat{S})]. \quad (4)$$

where $\hat{S}$ is the softmax applied to the result S_t. In order to reduce the value of *c* in the proposed method, the learing algorithm used Adam [3] optimizer. The learning rate is 0.0001 for 90K times. UCF101 [4] dataset was used for Training, a total of 9530 videos of 101 video classes were trained in 112x112 size. The structure of the our method is shown in Figure 1.

III. Experiment Results

For the performance evaluation of the proposed method, we experimented with Intel Core i7-7700K (4.20GHz) CPU, 4 Core, 8GB Ram and NVIDIA GeForce GTX 1080Ti system. As a database for experimental images, 101 video classes of UCF101 and 3790 test videos were used.

Table I shows an ablation study of the accuracy of the proposed method. In conclusion, the proposed method im proves 7.57% compared with RGB image only and 4.78% when moving average is applied.

Table II compares the proposed method with other papers. When tested with the same dataset, the performance is 7.57% higher than the based-line and 5.07% higher than Ishan et al.

IV. Conclusion

To recognize the actions we proposed a deep feature map that contains both spatial and temporal information of the input image. The proposed method combines spatial and temporal information using RGB image and differential image. In addition, recognition accuracy is increased by weighting important frames through moving average. As a result of experiment, we can get more improved performance when using the proposed deep feature map. Therefore, the proposed approach can be applied to intelligent monitoring system and automation system with quick response through action recognition and real-time safety-related action identification.

TABLE I: Ablation Study

Model	Accuracy(%)
RGB	48.4
RGB + Diff	54.62
RGB + Diff + Moving avg.	55.97

TABLE II: Accuracy comparison

Model	Accuracy(%)
Based-line [1]	48.4
Ishan et al [2]	50.9
The proposed method	55.97

Acknowledgment

This research was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-00250, and by In telligent Defense Boundary Surveillance Technology Using Collaborative Reinforced Learning of Embedded Edge Camera and Image Analysis).

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. [2] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: unsupervised learning using temporal order verification,” in *European Conference on Computer Vision*. Springer, 2016, pp. 527–544.

[3] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[4] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.