Effective Indoor Fire Detection with Channel Shuffle Module

Haotian Ge
School of Automation Southeast University; and Key Laboratory of Measurement and Control of Complex Systems of Engineering Ministry of Education, Nanjing, China
gehaotian96@163.com

Yichao Cao
School of Automation Southeast University; and Key Laboratory of Measurement and Control of Complex Systems of Engineering Ministry of Education, Nanjing, China
caoyichao@seu.edu.cn

Xiaobo Lu*
School of Automation Southeast University; and Key Laboratory of Measurement and Control of Complex Systems of Engineering Ministry of Education, Nanjing, China
xblu2013@126.com

ABSTRACT

In recent years, methods based on computer vision and deep learning become the mainstream approaches in fire detection. However, the expensive computation cost of 3D convolutional neutral network (CNN) is unbearable and it is difficult for them to capture the fire regions of videos in time. In this paper, we design a module named channel shuffle module (CSM) based on 2D CNN to keep the balance between computation cost and accuracy. By fusing RGB frame and differential frame, CSM improves the ability of 2D CNN in temporal information extraction which much less cost than methods based on 3D CNN. Four different structures of CSM are proposed and we choose the best one by experiment results. Also, experiments prove that the performances of TSN and TSM are improved with CSM in sequence classification. The accuracy of TSM with CSM is 99.2045%, false positive rate reaches 0.7890% and false negative rate reaches 0.4530%, which demonstrates the efficiency of CSM in temporal feature modeling.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Image and video acquisition**;

KEYWORDS

Fire detection, Sequence classification, Temporal information extraction, Channel shuffle

ACM Reference Format:
Haotian Ge, Yichao Cao, and Xiaobo Lu*. 2021. Effective Indoor Fire Detection with Channel Shuffle Module. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487125

1 INTRODUCTION

Fire is a great threat to public safety and social development, it often occurred in indoor places like residential buildings, enterprises and commercial centers. For a long a time, people have been relying on a variety of sensor equipment for indoor fire recognition. However, the small monitoring range and high maintenance cost are unacceptable. In recent years, with the rapid development of computer vision, the method based on deep learning provides a new solution for fire recognition. The process of fire recognition can be viewed as the classification for fixed length video sequences. It is easy to achieve the balance between accuracy and efficiency by using deep learning models.

There are mainly two solutions for video sequence classification: deep networks based on 3D convolutional neural network (CNN) and deep networks based on improved 2D convolutional neural network (CNN). Conventional 2D CNNs achieve excellent performance on the classification of individual frames. However, the biggest difference between video sequence frames and single frame is that video sequence frames are rich in temporal information. Traditional 2D CNNs cannot model the temporal information since all the parameters are used to model the spatial information, but 3D CNNs can make up for it. Tran et al. [1] expanded 2D convolution layer to 3D convolution based on VGG models named C3D. This network can learn the spatial and temporal information simultaneously from video sequence. Carreira and Zisserman [2] proposed I3D which inflate 2D filters and pooling kernels into 3D, so the network can extract spatio-temporal features. P3D [3] chooses to decompose 3D convolution operation into 1x3x3 2D convolution operation on spatial domain and 3x1x1 2D convolution operation on temporal convolution. Feichtenhofer et al. [4] presented SlowFast network. This network has two streams: Slow pathway and Fast pathway, which are used to capture spatial semantics and temporal semantics separately.

Although 3D CNNs perform well on the accuracy of the video sequence recognition, the heavy computation makes them difficult to meet the requirement of real-time recognition and judgement. So there have been various attempts to make 2D CNNs have the ability to extract more temporal information [5–7]. Wang et al. [8] designed TSN for action recognition in video. TSN uses sparse sampling strategy over long-range video sequence and extract aggregated features on sampled frames by 2D CNN. Zhou et al. [9] proposed TRN to strengthen the capacity to reason the temporal relations between frames. Lin et al. [10] presented a module named TSM to help 2D CNNs to perform as well as 3D CNNs with a little price. It can extract effective information of neighboring frames by shifting channels. ECO [11] uses long-term content already computed in the network to improve the performance of 2D CNNs on action recognition. In addition, some works [12–14] choose to mix the 2D and 3D convolution to capture useful information among the video sequence frames.

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487125

In this paper, we present a pre-processing method to enhance the temporal information of input image by inserting the Channel Shuffle Module (CSM) before the network. Since the huge parameters and calculation makes 3D CNNs hard to be applied, TSN and TSM are chosen as the base model for fire recognition algorithm. However, these base models often miss tiny fire objects in video sequences due to the lack of temporal features. CSM uses the composed frame fused by raw frame and differential frame from neighboring frames to increase the dynamic information for input images. CMS can be designed in 4 ways and the experiments will be conducted on these different structures.

The remaining parts of the paper are organized as follows. Section II will briefly introduce the frameworks of TSN and TSM. The design of CSM will be explained in section III. Experiments and the analysis of results are discussed in section IV. The conclusion is given in section V.

2 RELATED WORK

2.1 Temporal Segment Network (TSN)

Before the TSN was put forward, two-stream CNN was selected as the general solution for video sequence classification algorithm based on 2D CNN. One way in two-stream CNN is used to extract spatial feature from RGB images, and the other way is used to extract temporal feature from optical flow images or RGB differential images. But it can only capture short-term information in temporal domain from neighboring frames. In order to solve the problem mentioned above, TSN adopts sparse sampling strategy and segmental consensus function to enable the network to model the dynamic information over video sequence.

For fire recognition algorithm, if one of the frames in sampled sequence is judged to contain fire region, the whole sequence should be predicted to belong to the class of fire with higher probability. Therefore, we choose weighted averaging as the aggregation function. Weighted averaging function is derived as:

$$\sum\nolimits_{j=1}^{K} \omega_{ij} = 1 \tag{1}$$

$$\omega_{ij} = C_i\left(F_j\right)\left(\sum\nolimits_{m=1}^{K} C_i\left(F_m\right)\right)^{-1} \tag{2}$$

where ω_{ij} is the weight of score for input frame F_j in class i. The higher score should be given greater weight to ensure that it can contribute more to the final prediction in current category. It is beneficial for the network to determine the correct class over the video sequence with part of snippets containing fire.

3 OUR METHOD

3.1 Channel Shuffle Module (CSM)

The size of filter determines that 2D CNNs cannot extract temporal feature from higher dimension. The efficient way to improve the performance of 2D CNNs on video sequence classification is to stack the temporal information on one of the dimensions in images. Since the RGB images are rich in spatial feature and differential images are rich in temporal feature, if two kinds of images can be fused in some way, the new image will contain spatial and temporal information simultaneously. Inspired by this point, we proposed CSM as the pre-processing method to enhance the dynamic information of input images. The structure of CSM combined with TSN is present in Figure 1 (a) and the operation of CSM is illustrated in Figure 1 (b). Concretely, the frame i in sampled sequence $\{F_1, F_2, \cdot s, F_k\}$ is F_i, the differential frame generated by frame i and frame $i + 1$ is D_i. D_i can be computed in two ways. The first way is shown as:

$$\begin{cases} R\left(D_i\right) = R\left(F_{i+1}\right) - R\left(F_i\right) \\ G\left(D_i\right) = G\left(F_{i+1}\right) - G\left(F_i\right) \\ B\left(D_i\right) = B\left(F_{i+1}\right) - B\left(F_i\right) \end{cases} \tag{3}$$

where $R(\cdot)$ represents the red component of image, $G(\cdot)$ and $B(\cdot)$ represents the green and blue component separately. The differential frame is the result of the subtraction in corresponding channel components for neighboring frames. The second way is defined as:

$$D_i = gray\left(F_{i+1}\right) - gray\left(F_i\right) \tag{4}$$

where $gray(\cdot)$ represents gray processing for RGB image. The differential frame is computed by the subtraction between gray images converted from neighboring frames.

The operation of CSM is to shuffle the RGB image and differential image on the dimension of channel, so the fused image will contain temporal features without the increase of dimension. It means that we do not need to change the structure of current popular 2D CNNs for video sequence recognition because the dynamic information is added to the input images. Formally, the operation of CSM is:

$$N_i = F_i \odot D_i \tag{5}$$

where N_i is the fused image and $\odot$ is the specific shuffle operation in CSM. If more information from neighboring frames is added to the input image, the capacity of network in temporal domain modeling will be enhanced.

There are 4 approaches to shuffle the RGB image and differential image for CMS, as is shown in Figure 2. Type I chooses to split the RGB image and differential image in channel dimension. Both are divided into three components: red, blue and green, so the computation method of differential image is chosen as equation 3). The red component of differential image is inserted between red component and green component of RGB image. The blue and green component of differential image are inserted in a similar way. In Type II, the computation method of differential image is changed to equation 4), so the size of channel dimension in differential image is 1. We extent the channel dimension to 3 by copying the differential image and insert them like Type I. Type III and Type IV choose to stack the differential image after the green dimension of RGB image. The only difference of them is the computation method of differential image: Type III uses equation 3) while Type IV uses equation 4). The results of experiment for these 4 methods will be present in next section.

4 EXPERIMENT

4.1 Datasets and Evaluation Methods

All the frame sequences in datasets are cropped from monitoring video collected by ourselves. To increase the differences of pixel value, we choose different sampling strategies for videos with different frame rates. Concretely, there are 2 kinds of frame rate for monitoring video: 12 and 25, the sampling intervals are selected as 2 and 4 separately. The length of frame sequence is 32 and the size of each image is 224×224. The methods of data enhancement like

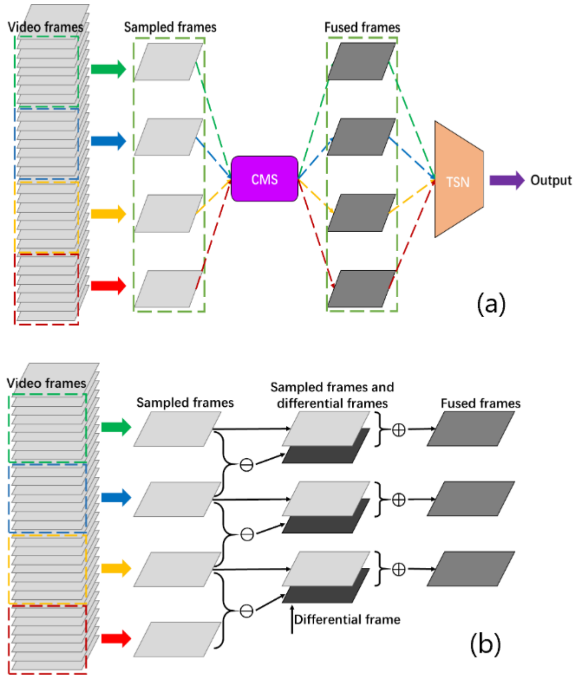

Figure 1: (a) The Structure of CSM Combined with TSN. (b) The Operation of CSM.

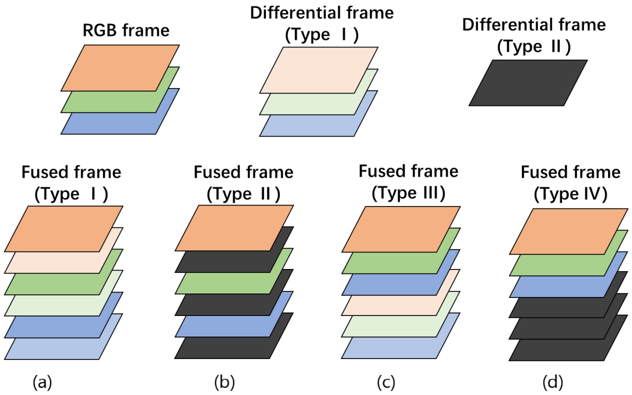

Figure 2: The Operation of Channel Shuffle in CSM. (a) Type I. (b) Type II. (c) Type III. (d) Type IV.

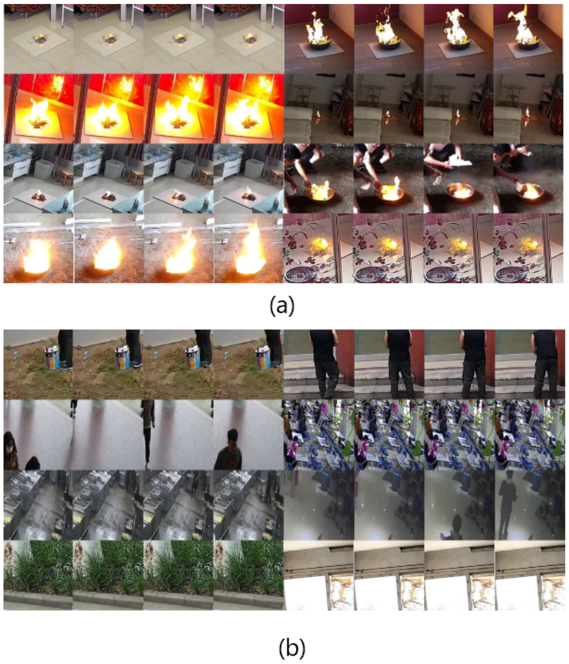

Figure 3: (a) Positive Samples. (b) Negative Samples.

resizing, random cropping, horizontal clipping will be carried out before the forward inference of the network.

The process of fire recognition can be converted to video sequence classification. The frame sequences are divided into 2 categories: sequences with fire region (positive samples) and sequences without fire region (negative samples). There are 9600 positive samples and 11480 negative samples in training set, 2535 positive samples and 2870 negative samples in testing set. In over 15% positive samples, the area of fire region is less than 30 pixels. Part of positive samples and negative samples are shown in Figure 3

In order to evaluate the performance of different networks on the datasets, we adopt accuracy rate (AR), false positive rate (FPR) and false negative rate (FNR) as evaluation criteria:

$$AR = (TP + TN)/(TP + FP + TN + FN) \tag{6}$$

$$FPR = FP/(FP + TN) \tag{7}$$

$$FNR = FN/(FN + TP) \tag{8}$$

where TP means true positive. TN means true negative. FP means false positive and FN means false negative. AR represents the performance of network in correctly recognizing different objects. FNR is used to evaluate the ability of network in capturing the sequences with fire region while FPR is used to evaluate the ability in the recognition of background objects. For fire recognition algorithm, we hope the network has high value in AR and lower values in FPR and FNR. Relatively speaking, FNR is more important than FPR in fire recognition, since the loss of missing fire target is much higher than false alarm in application.

4.2 Results

We design 4 structures for CSM to fuse the RGB image and differential image. The basic network is TSN and the backbone of it is ResNet-50. Our results are shown in Table 1

As is shown in Table 1, all the structures of CSM achieve better performance compared with TSN in RGB input on AR and FPR, but FNR of CSM in Type I and Type II are lower than TSN in RGB input. Only CSM in Type III and Type IV performs better than TSN in RGB and differential input and CSM in Type III achieves the best

Table 1: The Performance of CSM in Different Structures on Testing Set

Model	AR	FPR	FNR
TSN(RGB)	94.97%	6.27%	3.94%
TSN(RGB+RGBdiff)	96.48%	4.26%	2.86%
TSN+CSM(Type I)	95.41%	4.54%	4.63%
TSN+CSM(Type II)	95.73%	4.46%	4.11%
TSN+CSM(Type III)	**97.45%**	**2.92%**	**2.23%**
TSN+CSM(Type IV)	96.67%	3.43%	3.24%

Table 2: The Performance of Different Models Equipped with CSM

Model	AR	FPR	FNR
TSN(RGB)	94.97%	6.27%	3.94%
TSN+CSM(Type III)	97.45%	2.92%	2.23%
TSM(RGB)	98.76%	1.62%	0.91%
TSM+CSM(Type III)	**99.20%**	**0.79%**	**0.45%**

performance in all these criteria. The computation of CSM in Type III is the same with other types. Also, the speed of TSN with CSM in Type III in forward inference is 1.3 times that of TSN in RGB input, while the speed of TSN with two-stream input in forward inference is 3 times than the speed of TSN in single input. It proves the effectiveness of our CSM with little extra computation. We choose Type III as the structure of CSM in following experiments. Table 2 shows the performance of TSN and TSM equipped without and with CSM on all the criteria. The backbones of TSN and TSM are both ResNet-50 and CSM is chosen in Type III.

The results in Table 2 prove that the performance of TSN and TSM is much better in these criteria after equipping with CSM. CSM enhances temporal modelling ability of raw networks further based on their original performance. This makes us confirm that CSM has the ability of generalization is strong when it is added to different models.

In video test, we use FPN [15] as the basic model of detection algorithm to capture the suspected fire regions in the full frame of video. All the suspected regions will be classified by TSM with CSM in Type III. The suspected region captured by FPN is marked by blue bounding box and red number represents confidence probability. The fire region confirmed by TSM with CSM is marked by green bounding box and the size of green bounding box is 224×224. The testing results of TSM with CSM on monitoring videos are present in Figure 4. All the images on the top do not contain fire region while images on the bottom are on the contrary in Figure 4. We can find that TSM with CSM can capture all the fire region in the frames correctly even if the area of flame is very small, which proves the effectiveness of CMS in fire recognition algorithm.

5 CONCLUSION

In this paper, a new module named CSM is proposed to improve the performance of models on fire detection by enhancing temporal modeling ability through fusing RGB image and differential image. Four different structures are designed for CSM and we choose the best one (Type III) from results of experiments. Then CSM in Type III is added to TSN and TSM, and experiments prove that the performances of models in AR, FPR and FNR are improved. The AR, FPR and FNR of TSM with CSM in Type III is 99.20%, 0.79% and 0.45% separately. Also, video test is conducted to show the performance of CSM on real surveillance videos. However, there is much room for improvement on different evaluation criteria and CSM does not utilize the information between frames from large intervals. Therefore, the future work will focus on these points to improve the capacity of models in temporal feature extraction.

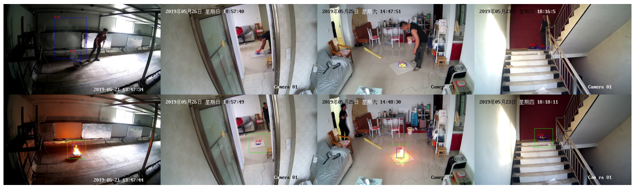

Figure 4: The Testing Results of TSM with CSM on Monitoring Videos Samples.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61871123), Key Research and Development Program in Jiangsu Province (No.BE2016739) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

[1] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. IEEE International Conference on Computer Vision. IEEE.
[2] Carreira J, Zisserman A (2017). Quo vadis, action recognition? a new model and the kinetics dataset. IEEE.
[3] Qiu Z, Yao T, Mei T (2017). Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE.
[4] Feichtenhofer C, Fan H, Malik J, K He (2019). SlowFast Networks for Video Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV).

IEEE.

[5] X Wang, Girshick R, Gupta A, He K (2018). Non-local Neural Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[6] K He, X Zhang, S Ren, and J Sun (2016). Deep residual learning for image recognition. 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[7] Szegedy C, Wei L, Jia Y, Sermanet P, Rabinovich A (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[8] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, *et al.* (2016). Temporal segment networks: towards good practices for deep action recognition. European Conference on Computer Vision.

[9] Zhou B, Andonian A, Oliva A, Torralba A (2018). Temporal relational reasoning in videos. European Conference on Computer Vision.

[10] Lin J, Gan C, Han S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.

[11] Zolfaghari M, Singh K, Brox T (2018). Eco: efficient convolutional network for online video understanding. European Conference on Computer Vision.

[12] Du T, Wang H, Torresani L, Ray J, Paluri M (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[13] Xie S, Sun C, Huang J, Tu Z, Murphy K (2017). Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification.

[14] Sun L, Jia K, Yeung D, Shi B (2015). Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks. IEEE International Conference on Computer Vision pp 4597-4605. IEEE.

[15] Lin T, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017). Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society.