Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 493–498

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop Internet of Smart Things (IST 2021)
November 1-4, 2021, Leuven, Belgium

Video-based Fire Smoke Detection Using Temporal-spatial Saliency Features

Zili Zhang[a,b], Qingyong Jin[a,b], Lina wang[a,b], Zhiguo Liu[a,b,*]

[a]*Department of Computer Science and Engineering, Shijiazhuang University, No.288, Zhufeng Street, Shijiazhuang, 050035, China*
[b]*Hebei Province Internet of Things Intelligent Perception and Application Technology Innovation Center, No.288, Zhufeng Street, Shijiazhuang, 050035, China*

Abstract

In this paper, we propose a video based spatial-temporal convolutional neural network for fire smoke recognition. The model concatenates the appearance features and the motion features followed by a convolution layer to implement spatial-temporal feature fusion. To reduce the influence of background of no-smoke, we use an attention module to capture salience features from the input image. Experiments on the self-created dataset show that the presented method is valid, which achieves a detection rate of 97.5% and accuracy rate of 96.8%.

Keywords: Fire detection; Temporal-spatial features; Video;

1. Introduction

Early fire detection is important for reducing losses caused by fire as the damage often increases exponentially over time. Fire detection based on surveillance video is a promising detection method especially in open or large spaces and outdoor environments. Compared to traditional detection methods generally based on multiple sensors [1, 2], video fire detection systems have many advantages [3]. However, it is still a challenging task due to various appearance of smoke. Gaur et al. [4] give a review about video flame and smoke based fire detection methods. Existing fire detection algorithms based on video can be divided into two types, the method based on traditional handcrafted features and the approach based on Convolutional Neural Network (ConvNet). For traditional smoke detection algorithms [5, 6, 7], it is difficult to achieve high detection rate as it depends on handcrafted features selected mainly empirical. Compared with the handcrafted features based method, ConvNet can automatically learn a unique set of features for give tasks.

* Corresponding author. Tel.: +86-311-66614195 ; fax: +86-311-66614195.
E-mail address: sjzlzg@163.com

1877-0509
10.1016/j.procs.2021.12.275

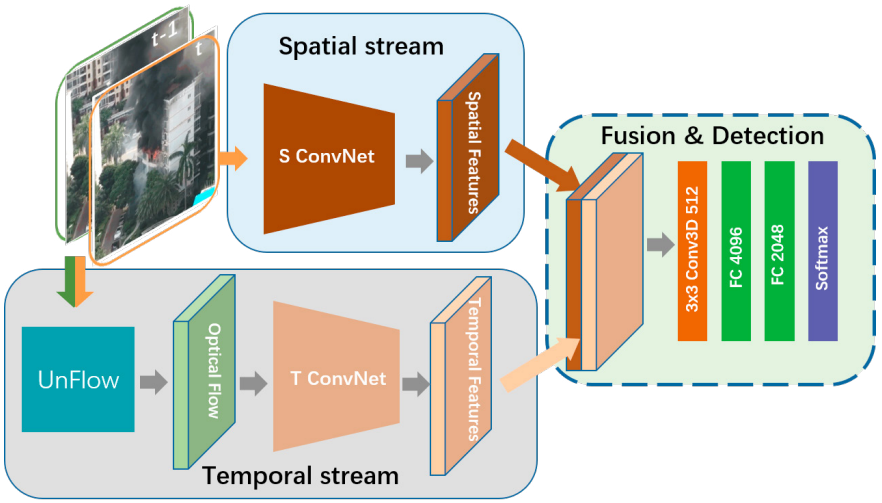

Fig. 1. The presented model contains of a spatial stream, a temporal stream and a fusion-detection module. The spatial stream takes a video frame as input and captures the appearance features. The temporal stream takes two neighbor video frames as input and learns the motion features. The fusion-detection module outputs the detection results according to the spatial-temporal fusion features.

ConvNet architecture has significantly pushed the performance of visual processing tasks [8, 9, 10, 11] based on their rich representation power.

Inspiring by the great success of ConvNet in vision tasks, many fire smoke detection methods based on ConvNet are proposed [12, 13, 14]. Tao et al. [12] proposed a novel method based on deep convolutional neural networks to extract smoke features from image. To utilize the motion information between frames, Hu and Lu [13] presented a spatial-temporal architecture. Lin [14] proposed a novel model based on 3D CNN for recongnizing smoke by combining dynamic spatial-temporal information.

These methods use the features of input image to recongnize fire smoke, however, smoke usually is a small part of the input image, which causes detection error due to the influence of the background. We propose a spatial-temporal ConvNet which fuses the appearance features and the motion features. Meanwhile, we utilize an attention module to extract the salient parts of spatial and temporal feature maps. The main contribution is as follows:

- We propose a novel end-to-end convolutional neural network combing the spatial and motion features to detect fire from the input video.
- We combine an attention module to avoid the influence of noise and improve the detection accuracy.
- We validate the effectiveness of the proposed model by testing it on different videos.

2. Method

In this section, we describe the proposed model. We decompose video sequence into two components, i.e., spatial component and temporal component. The proposed model consists of two streams, i.e., spatial stream and temporal stream, to capture the two components. As the fire smoke area is usually a small part of monitoring video images, focusing the important features and suppressing unnecessary ones can obtain improvement of detection accuracy. Therefore, we insert an attention module CBAM to refine the raw spatial features and motion features.

2.1. Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) [15] is an effective attention module for feed-forward Convolutional neural networks. As shown in Fig. 2, the input of the model is a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. Then the model

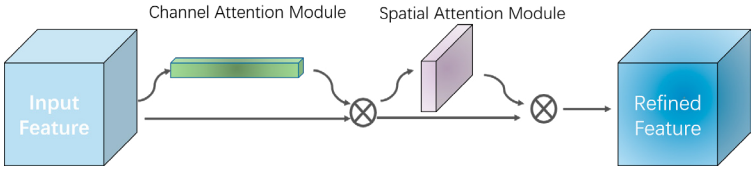

Fig. 2. The CBAM contains a channel attention module and a spatial attention module.

generates attention maps from both the channel and spatial dimensions. Finally, the attention maps are multiplied to the input feature map for adaptive feature refinement.

In short, the overall attention process is described as follows:

$$\begin{aligned} \mathbf{F}^{'} &= M_c(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F}^{''} &= M_s(\mathbf{F}^{'}) \otimes \mathbf{F}^{'} \end{aligned} \tag{1}$$

where $\otimes$ denotes element-wise multiplication. $M_c(\cdot)$ and $M_s(\cdot)$ are the processes of channel attention and spatial attention, respectively.

Channel attention focuses on the meaningful part of an image. First, two different spatial context descriptors, $\mathbf{F}^c_{avg} \in \mathbb{R}^{C\times1\times1}$ and $\mathbf{F}^c_{max} \in \mathbb{R}^{C\times1\times1}$, are obtained by applying global average-pooling and global max-pooling on the input feature map $\mathbf{F}$. Then, both descriptors are forwarded to a multi-layer perception (MLP). Finally, the channel attention map $\mathbf{M}_c \in \mathbb{R}^{c\times1\times1}$ is obtained by merging the output feature vectors of the MLP using element-wise summation. In short, the channel attention map is computed as:

$$M_c = \sigma\left(MLP\left(AvgPool\left(\mathbf{F}\right)\right) + MLP\left(MaxPool\left(\mathbf{F}\right)\right)\right) \tag{2}$$

where σ denotes the sigmoid function.

Different from the channel attention, spatial attention focuses on the location of an informative part. First, two 2D maps, $\mathbf{F}^s_{max} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F}^s_{avg} \in \mathbb{R}^{1\times H\times W}$, are generated by using max pooling and average pooling on the input feature map $\mathbf{F}^{'}$ for aggerating channel information. Then the two feature maps are concatenated. Finally, the concatenated result is convolved by a standard convolution layer to produce the 2D spatial attention map. Therefore, the spatial attention is calculated as:

$$M_s = \sigma\left(f^{7\times7}\left(\left[AvgPool\left(\mathbf{F}\right); MaxPool\left(\mathbf{F}\right)\right]\right)\right) \tag{3}$$

where $f^{7\times7}$ denotes a convolution operation with the filter size of 7×7.

2.2. Network Structure

The proposed model is composed of the spatial stream, the temporal stream and the fusion-detection module as shown in Fig. 1. Each stream is implemented using a deep ConvNet and performs feed forward respectively. The input of the fusion-detection module is the spatial-temporal feature map concatenating both the spatial features and the temporal features.

Spatial Stream The spatial branch uses a video frame as input. It mainly learns the appearance features of fire smoke, such as color, shape and texture, for fire smoke detection. The architecture of the stream is shown in Fig. 3. It

contains five convolutional layers, four pooling and a CBAM. The first four convolution layers are followed by max pooling. The attention module, CBAM, which can emphasis the fire smoke features, follows after the output of the forth convolutional layer. The input video frame is a square RGB image of $227 \times 227 \times 3$ and the first convolutional layer filters the input with 96 kernels of $7 \times 7 \times 3$. The second convolution layer take the output of the first layer and filters it by using 256 kernels of size $3 \times 3 \times 96$. Note that the size of kernels for other layres are shown in Fig. 3. The input of the CBAM module is the feature map of $47 \times 47 \times 256$. Finally, the spatial stream outputs the deep feature maps of $37 \times 37 \times 256$ denoted as $\mathbf{F}_S$.

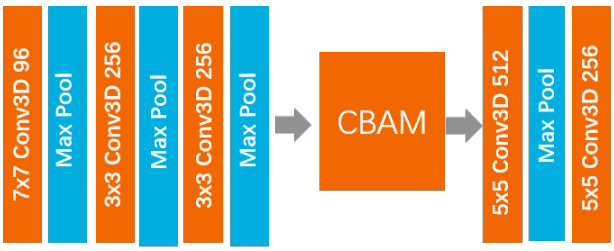

Fig. 3. The architecture of the spatial stream, S. The CBAM module follows after the third convolution layer for refining the spatial features.

Temproal stream In contrast with spatial stream, we concatenate neighbor two frames of a video sequence as the input of the temporal stream. To describe the temporal features, we estimate the optical flow displacement fields between two consecutive frames as motion features. As obtaining dense perpixel optical flow is difficult for real scenes, we use the unsurvised flow estimatino network UnFlow to generate optical flow for the input video sequence. The architecture of the temporal stream is similar to the spatial stream. It mainly contains the UnFlow moudule, several convolutional layers and the CBAM. The size of optical flow estimation is $227 \times 227 \times 2$ which represents the horizontal component and vertical component, respectively. Some convolutional layers follow the UnFlow module to learn the motion features. The first convolutional layer takes as input the output of the module UnFlow and filters it filters with 96 kernels of size $7 \times 7 \times 2$. The parameters of other layers are show in Fig. 4. The temporal stream outputs the deep feature maps of $37 \times 37 \times 256$ denoted as $\mathbf{F}_m$.

Fig. 4. The architecture of the temporal stream, T. The module UnFlow is used to learn the dense optical flow. The CBAM module is used to capture important motion features.

Fusion and Detection The fusion-detection module recognizes fire or non-fire by fusing the spatial features and the temporal features. We concatenate the two features, $\mathbf{F}_s$ and $\mathbf{F}_m$, to generate the fusion feature map as the input of the fusion-detection module. The size of fusion feature map is $37 \times 37 \times 512$ denoted as $\mathbf{F}_f$.The module contains a convolutional layer, two fully connected layers and a softmax classifier. The fusion feature is first filtered by the convolutional layer with 512 kernels of size $3 \times 3 \times 512$. The softmax layer output the detection result for the input video frames.

2.3. Implemention

We define the training set $\mathbf{X} = \{x^1, x^2, X^3, \cdots, x^N\}$, $x^i = (I_i^0, I_i^1, l^i))$, where N denotes the amount of the dataset, x^i is the i-th training sample, l^i represents its class label (fire or non-fire), I_i^0, I_i^1 are two consecutive neighbor video frames. The UnFlow firstly are trained by the fire smoke dataset. Then, we utilize multi-task learning strategy to train

the two-stream model. At the training stage, both the spatial stream and the temporal stream are trained in turns and optimally reach an equilibrium. To train the proposed network, we design the loss function as followed:

$$\mathcal{L} = \underset{a}{arg\ min}\, E(-\sum_{i=1}^{K} \delta(i = l^i) log\ P(i = l^i | [S(I_i^1, \theta_s); T(I_i^0, I_i^1 \theta_t)], \theta_f)) \tag{4}$$

where $\delta(.)$ is the indicator function. $P(.|.)$ represents the predicted probability distribution. $S(.)$ and $T(.)$ are the processes of the spatial stream and the temporal stream, respectively. θ_s denotes the parameters of the three components, S, T, $F\&D$, respectively.

The network weights are learnt using the mini-batch stochastic gradient descent with momentum (set to 0.9). At each iteration, a mini-batch of 128 samples is constructed by sampling from our dataset.

3. Experiments

This section reports the experimental validation of our approach. We use the open source toolbox PyTorch to implement the proposed module. All networks are trained on a 12GB NVIDIA Titan X GPU.

Dataset Existing fire detection dataset lack some hard negative samples, therefore, we create our own training set consisting of 116 fire video and 89 non-fire videos. We collect fire smoke data from Youtube, some open dataset and take fire videos through some fir drills. The dataset has some challenging videos such as cloud and fog. We truncate these videos to 5328 clips which include 2836 positive samples and 2492 negative samples. Plentiful high-quality data is the key to improve the performance of model. To avoid over-fitting, we adopt horizontal reflection and rotation to augment dataset. The dataset is split into 80% and 20% between non-overlapping training and validation subsets.

Quantitative Analysis In this section, we show the experiment results of our method and comparisons with other methods. To quantitatively analysis the experiment results, we take Accuracy Rate (AR) and Detection Rate (DR) as the evaluation criteria.

The method [12] is based on AlexNet [16] and takes a full image as full. It classifies smoke or non-smoke by extracting appearance features, since the module can be regarded as a single spatial stream ConvNet. We test the method in our dataset. The AR and DR are 93.8% and 94.7%, respectively. Hu and Lu [13] proposed an enhanced spatial-temporal convolutional neural network which captures spatial features and motion features. It utilizes a multi-task learning strategy to jointly recognize smoke and estimate dense optical flow. We repeated their implementations in our dataset and achieve the AR of 96.3% and DR of 97.2%. 3D Convolutional nerual networks can be successfully used for video analysis. Lin et al. [14] adopt 3D ConvNet to realize smoke detection by combing dynamic spatial-temporal information. We also repeated the method in our dataset and achieve the AR of 96.5% and DR of 96.9%. Our method utilizes two-stream strategy and saliency analysis to detect fire smoke. It improves the detection ability and achieve the state-of-art performance with the AR of 96.8% and DR of 97.5%.

Table 1. Results of our approach and other deep learning based methods

Methods	AR	DR
Spatial stream ConvNet	93.8%	94.7%
Enhanced spatial-temporal ConvNet	96.3%	97.2%
3D ConvNet	96.5%	96.9%
Ours	96.8%	97.5%

Ablation Analysis In this experiment, we study the effect of the attention module CBAM. First, we remove the CBAM module from both the spatial stream and the temporal stream. Then we train the new network in our dataset and test it on the validation subsets. Finally, both the Accurancy Rate and Detection Rate are 96.1% and 96.5%,

respectively. Comparing with the results of the two-stream architecture with CBAM, the attention module can improve the accuracy of recognize fire smoke.

To evaluate the temporal stream, we remove it from the proposed architecture and utilize Accuracy Rate and Detection Rate as evaluation criteria. We test it on the validation subsets, obtaining the AR of 93.7% and DR of 94.2%, which is worse the two-stream architecture.

4. Conclusions

In this paper, we presented a spatial-temporal convolutional neural network for detecting fire smoke from video. The spatial stream focuses on learning the appearance information and the temporal stream mainly captures the motion features. The results show that our approach achieves highest detection rate. In the future, we would like to try designing a lightweight deep learning model for fire detection. Meanwhile, improving the accuracy of motion estimation maybe another potential research interest.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This paper is supported by the Science and Technology Research and Development Program of Shijiazhuang (No. 211130203A) and the project of Hebei University Science and technology research project (ZD2021421).

References

[1] Osman S da Penha and Eduardo F Nakamura. Fusing light and temperature data for fire detection. In *The IEEE symposium on Computers and Communications*, pages 107–112. IEEE, 2010.

[2] Arnoldo Díaz-Ramírez, Luis A Tafoya, Jorge A Atempa, and Pedro Mejía-Alvarez. Wireless sensor networks and fusion information methods for forest fire detection. *Procedia Technology*, 3:69–79, 2012.

[3] A Enis Çetin, Kosmas Dimitropoulos, Benedict Gouverneur, Nikos Grammalidis, Osman Günay, Y Hakan Habiboğlu, B Uğur Töreyin, and Steven Verstockt. Video fire detection–review. *Digital Signal Processing*, 23(6):1827–1843, 2013.

[4] Anshul Gaur, Abhishek Singh, Anuj Kumar, Ashok Kumar, and Kamal Kapoor. Video flame and smoke based fire detection algorithms: A literature review. *Fire technology*, 56(5):1943–1980, 2020.

[5] B Uğur Töreyin, Yiğithan Dedeoğlu, and A Enis Cetin. Wavelet based real-time smoke detection in video. In *2005 13th European signal processing conference*, pages 1–4. IEEE, 2005.

[6] Mehdi Torabnezhad, Ali Aghagolzadeh, et al. Visible and ir image fusion algorithm for short range smoke detection. In *2013 First RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)*, pages 38–42. IEEE, 2013.

[7] Feiniu Yuan, Zhijun Fang, Shiqian Wu, Yong Yang, and Yuming Fang. Real-time image smoke detection using staircase searching-based dual threshold adaboost and dynamic analysis. *IET Image Processing*, 9(10):849–856, 2015.

[8] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[10] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021.

[11] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021.

[12] Chongyuan Tao, Jian Zhang, and Pan Wang. Smoke detection based on deep convolutional neural networks. In *2016 International conference on industrial informatics-computing technology, intelligent technology, industrial information integration (ICIICII)*, pages 150–153. IEEE, 2016.

[13] Yaocong Hu and Xiaobo Lu. Real-time video fire smoke detection by utilizing spatial-temporal convnet features. *Multimedia Tools and Applications*, 77(22):29283–29301, 2018.

[14] Gaohua Lin, Yongming Zhang, Gao Xu, and Qixing Zhang. Smoke detection on video sequences using 3d convolutional neural networks. *Fire Technology*, 55(5):1827–1847, 2019.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.