Multimed Tools Appl (2019) 78:237–256
https://doi.org/10.1007/s11042-017-5561-5

Recurrent convolutional network for video-based smoke detection

Mengxia Yin[1] · Congyan Lang[1] · Zun Li[1] ·
Songhe Feng[1] · Tao Wang[1]

Received: 30 July 2017 / Revised: 4 December 2017 / Accepted: 21 December 2017 /
Published online: 1 February 2018

Abstract Video-based smoke detection plays an important role in the fire detection community. Such interesting topic, however, always suffers from great challenge due to the large variances of smoke texture, shape and color in the real applications. To effectively exploiting the long-range motion context, we propose a novel video-based smoke detection method via Recurrent Neural Networks (RNNs). More concretely, the proposed method first captures the space and motion context information by using deep convolutional motion-space networks. Then a temporal pooling layer and RNNs are used to effectively train the smoke model. Finally, to promote further research and evaluation of video-based smoke models, we also construct a new large database of 3000 challenging smoke video clips that cover large variations in illuminance and weather conditions. Experimental results demonstrate that our proposed method is capable of achieving state-of-the-art performance on all public benchmarks.

Keywords Smoke detection · Motion context information · Deep convolution · RNNs

1 Introduction

Recently, smoke detection in video surveillance is a valuable technique in the early fire detection. It has attracted more and more researchers to make effort in this field. The goal of smoke detection is to prevent fires by detecting smoke at the early stage of fire, which is deemed as an essential problem in fire detection.

✉ Congyan Lang
cylang@bjtu.edu.cn

Mengxia Yin
18813172300@163.com

[1] Beijing Jiaotong University, Beijing, China

There are two main methods of smoke detection: video-based smoke detection and image-based smoke detection. Existing methods mainly combine motion region detection with various visual representation of smoke such as color, texture and shape features. In Fig. 1a shows the process of smoke detection methods based on moving region detection and feature extraction. In the early video-based smoke detection methods, color and motion characteristics are widely used to extract suspected moving areas. An energy-based smoke detection model is proposed in [32], properties in wave-let space are studied to detect smoke that are unstable to light. In [6, 8], a chrominance-based static decision rule and a diffusion-based dynamic decision rule with RGB contrast-image and shape constrain are proposed to reduce the interference of pure color objects in the wavelet domain. Millangarcia et al.[23] applies the rule to YCbCr color space and propose a new representing method. A computational intelligence classifier [11] is adopted to identify the presence of smoke combined with YUV color space. Krstini et al. [16] focuses on a pixel level analysis and segmentation of smoke colored pixels based on HSI color space. Being enlightened by the fact that smoke moves upward influenced by the heat, a fast accumulative motion orientation model [36] based on integral image is proposed. Together with the rule that R,G, and B values are close to each other, this algorithm has good efficiency and robustness.

With further study of human visual perception system and visual information mechanism, quantized visual features [3, 4] have been widely used in robust image analysis. Smoke texture [22, 28], diffuse and fuzzy characteristics [9] are also widely used in smoke

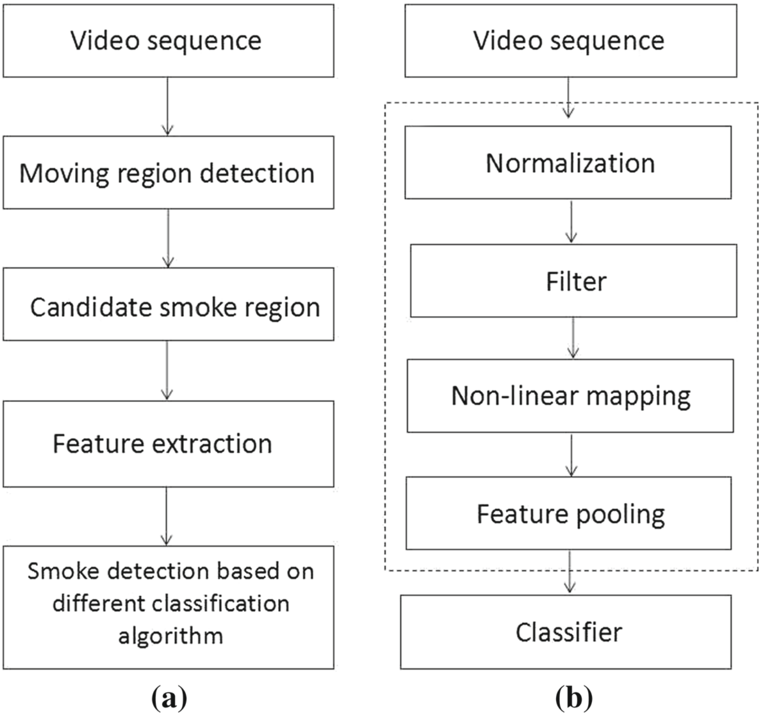


Fig. 1 The basic process of smoke detection: **a** shows the process of smoke detection methods based on moving region detection and feature extraction. Smoke detection based on Convolutional Neural Network (CNN) is described in (**b**)

detection. Additionally, to get high robustness and low false positive rate, the temporal and spatial characteristics are studied. Avgerinakis et al. [2] and Barmpoutis et al. [5] extract smoke candidate blocks through five continuous frames. Subsequently, histograms of oriented gradients (HOGs) and histograms of optical flows (HOFs) descriptors are used to distinguish moving smoke-color objects. To achieve intelligent smoke detection models, as a crucial part of artificial intelligence, machine learning is gradually applied to smoke detection system. Block based Inter-Frame Difference and dynamic texture features from smoke histogram image [7] are combined to train an SVM classifier. Later, local binary patterns (LBP) features of the candidate blocks [12, 37] are used to train AdaBoost. Combined with AdaBoost, a robust classifier [35] is constructed using multiple neural networks combined with BP to classify smoke and non-smoke objects. But it could not work well when smoke-color objects or transverse flow occur. A dual threshold AdaBoost algorithm with a staircase searching technique [38] is used for video smoke detection. More recently, a random forest classifier is built in the process of training by using bag-of-features (BOF) [25] which will make smoke detection near real-time and increase detection accuracy. Smoke detection of single images is still a challenging problem in both theoretical and practical implications. The recent advances [26, 27, 29] propose a novel feature to detect smoke of single image. An image formation model that regard an image as a linear combination of smoke and non-smoke components is proposed based on the atmospheric scattering models. The separation of the smoke and non-smoke components is formulated as convex optimization that solves a sparse representation problem. Using the separated quasi-smoke and quasi-background components, the feature is constructed as a concatenation of the respective sparse coefficients.

However, existing methods for smoke detection still have high false alarm and low detection rate. Traditional smoke detection methods based on feature extraction can not extract the characteristics of smoke accurately because they are vulnerable to light, airflow and obstruction. For example, the color of the smoke varies widely from white to light gray to black. Lots of non-smoke objects such as roads, gray clothes and clouds are similar to smoke in some extent. Additionally, the shapes and areas of smoke are greatly affected by the airflow. Furthermore, the clarity and range of smoke are susceptible to the occlusion of foreground objects resulting in unreliable features extracted from these blurred images. Thus, it is still of great challenge to detect smoke accurately.

Recently, deep learning methods have shown superior performance for many tasks such as image classification [21, 34], pedestrian detection [30, 31]and age estimation [1, 24]. Furthermore, these methods perform better in the fields of automatical makeup for female [18], face aging [19] and surveillance video parsing [20]. The multi-layer neural networks can learn representative and essential features of data and forms abstract high-level attributes through combining low-level features. It is reasonable to make smoke detection by multi-layer neural networks.

Taking the above issues into account, one needs to solve two key problems: i) extract effective features to represent the smoke images and ii) explore useful information to formulate the properties of the smoke regions over the extracted features. To achieve the goal, we propose a robust real-time smoke detection system which not only extracts effective features by motion and space convolution networks but also aggregates motion information by a recurrent method. Especially for learning motion context information, inspired by [17] which jointly learns appearance representation and motion context from adjacent frames using a two-stream convolutional architecture, we design a new deep convolutional motion-space networks to learn motion context information. Our method can achieve better performance compared to other smoke detection methods. Our contributions are as follows:

Firstly, we create a large-scale dataset that is more likely to real scenes. The smoke videos in our dataset cover real-world conditions, with large variations in illuminance and weather conditions, such as forest smoke, parking lot smoke and factory smoke. In addition, to improve the adaptability of smoke detection, the smoke video clips include the occlusion in a large degree. We call this dataset BJTU-Smoke. Secondly, in order to extract effective features to represent smoke, we propose a novel video-based smoke detection method by deep convolutional recurrent motion-space networks. In Fig. 1b shows the process of smoke detection methods based on CNN. In our model, there are two individual networks learning spatial representation and motion features from source video frames which are called space network and motion network. Finally, since the proposed model can learn both spatial representation and motion features from source video frames, we choose to fuse the corresponding features after the fully connected layer of the two networks. In addition, in order to obtain discriminative motion context information of a consecutive periods of time, the motion-space characteristics are aggregated in a recurrent way by RNNs which are enhanced by a temporal pooling layer to obtain the long-term context information of an entire sequence. Evaluation on our dataset shows the robustness and applicability of our method.

The paper is organized as follows. Section 2 describes the details of our datasets. In Section 3, the architecture of the proposed network is presented. Experimental results are demonstrated in Section 4, followed by conclusion drawn in Section 5.

2 Data construction

In this part, we will explain where and how we obtain the data and how we annotate it in detail.

2.1 Data collection

Existing widely used datasets such as the datasets from B.C. Ko [14], Toreyin et al. [32] and R. Vezzani [33] are relatively small. The largest dataset contains 20 videos from Toreyin et al. [32] with an average of no more than 3 minutes per video. The dataset from B.C. Ko [14] contains 16 video clips and there are 14 video clips in the dataset of R Vezzani [33]. These videos can not cover complex scenes. In order to simulate the real world scenes, smoke and non-smoke videos should be varied and abundant to ensure that the algorithm has a very good generalization. Therefore, we collect a large number of real-world videos by search engine using keywords from websites. Until now, our dataset covers smoke video clips from different cities, complex real-world scenes, and diverse objects. We take each frame from the raw videos. The diversity of our dataset helps to improve the robustness and adaptability of smoke detection.

2.2 Data annotation and statistics

The images are annotated manually next. Each smoke image is labeled 1 and non-smoke image is labeled 0 according to the traditional method. In other words, the images with smoke is positive samples but is negative samples without smoke. To be clear, the label is at the video frame level. In addition, being enlighten by [39, 40], we divide smoke images into two groups according to the proportion of smoke in an image, which are called small smoke with the proportion less than 20%. The small smoke images are shown in Fig. 2 and

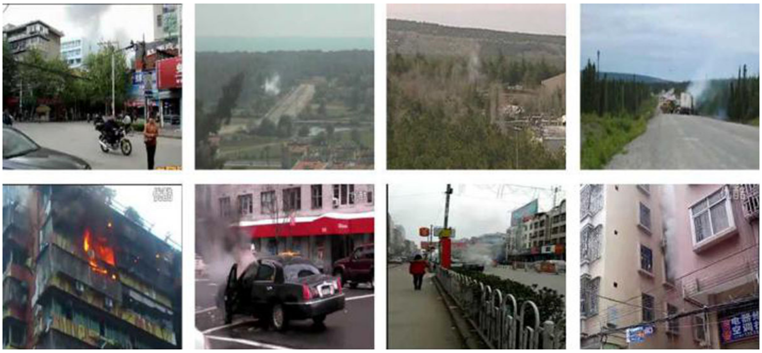


Fig. 2 Small smoke. Our dataset contains 1000 small smoke video clips with the smoke proportion of less than 20%

the rest are called large smoke that are shown in Fig. 3. The dataset contains a large number of images that are similar to smoke but are non-smoke in fact shown in Fig. 4.

Our new dataset contains 5000 video clips with about one minute per video after discarding clips with low resolution. There are 3000 smoke videos of these 5000 videos in total, the rest of which are nonsmoke videos. Although our videos cover different cities, complex scenes and diverse objects, an imbalance still exists between them. This is unavoidable because small smoke videos appear rarely.

In summary, our large-scale dataset falls into two classes, and there are many instances in each class. The images in the dataset have resolution 128 × 64 pixels and cover large variations in illuminance, weather and complex scenes. We use it to train our motion-space model for this purpose.

3 Deep convolutional recurrent motion-space networks

The purpose of deep learning is to discover a hierarchical model that can represent the data best. We propose a novel video-based smoke detection method by deep convolutional recurrent motion-space networks which consists of both space and motion network learning spatial representation and motion context information from source video frames. Firstly, we introduce the overall structure of the proposed network which is illustrated in Fig. 5. Next,

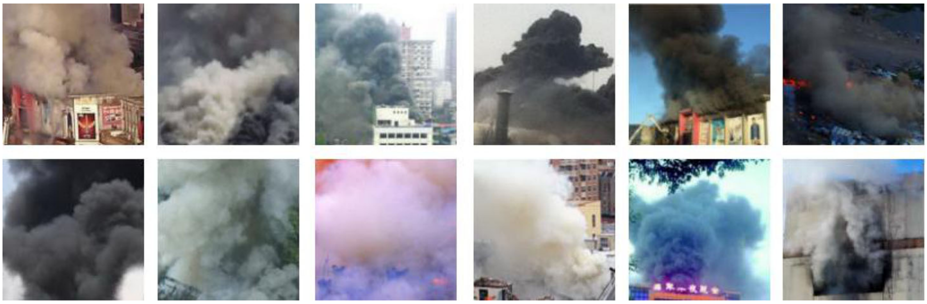


Fig. 3 Samples of large smoke: with the proportion of more than 20%. They cover smoke videos from different cities, complex real-world scenes, and diverse objects

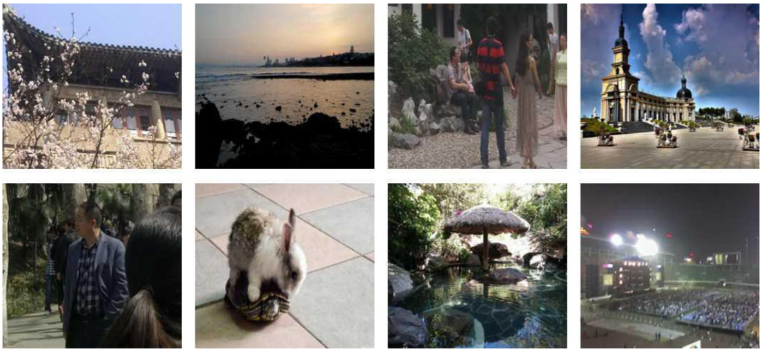


Fig. 4 Non-smoke samples. Non-smoke videos should be varied and abundant. A large number of real-world videos are collected from websites. The dataset contains a large number of images that are similar to smoke but are non-smoke in fact

for two consecutive smoke images, we describe the details of the two networks and how they work together. Finally, the fusion layers to fuse the motion and space information and fusion methods are stated.

3.1 Structure overview

Figure 5 illustrates the proposed deep convolutional recurrent motion-space network (RMSN). In this model, there are two individual networks learning spatial representation and motion features of each two consecutive frames from source video frames which are called motion network and space network. In detail, the space network (yellow rectangles) is used to learn spatial features from raw video frames while a pair of consecutive video frames of a smoke is processed by the motion network (blue rectangles) at each time-step to

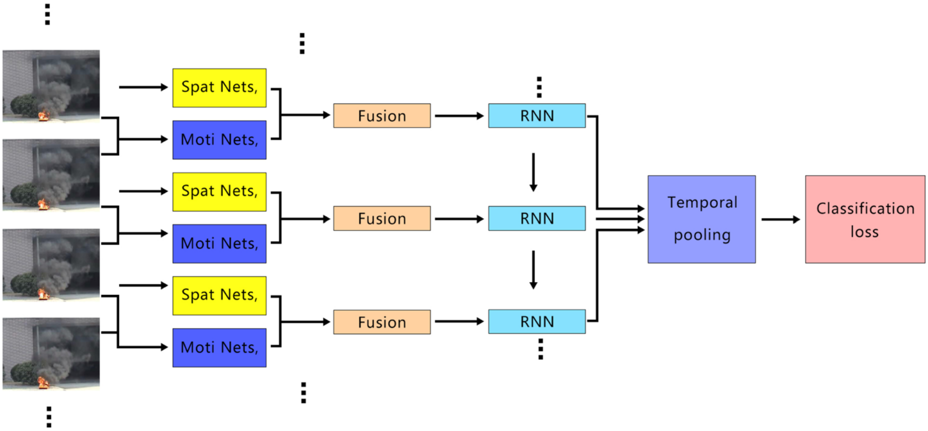


Fig. 5 The framework of our RMSN. Each two consecutive frames are processed by space network (yellow rectangles) and motion network (blue rectangles). Then these representative features are fused in a recurrent way for learning representative motion context information. These features are integrated by temporal pooling layer from any video sequence into a single feature representation. The whole RMSN network is trained by introducing soft-max loss function to detect smoke

predict motion between the adjacent frames. Later, this two individual networks features are fused in a recurrent way to learn discriminative and representative motion context information. A temporal pooling layer is used to obtain a single feature representation of any video sequence by integrating these features. Finally, for effectively training this whole model, we adopt the soft-max loss function as classification loss. The classification loss function predicts whether smoke is present in each video frame. Next, the implementation details of each components of the proposed model are introduced.

3.2 Deep convolutional motion-space networks

As described in 3.1, we know that the motion features and the space features are studied simultaneously by two convolution networks. Particularly, the spatial network is adopted to learn the spatial representation of the original video frames. The motion network is used to learn the motion characteristics of two consecutive video frames. In the following, we will introduce the detail architectures of the space and motion networks.

3.2.1 Motion network

As shown in Fig. 6, each two consecutive video frames of smoke are processed by the motion network in RMSN (corresponding to blue rectangles in Fig. 5) to capture and predict the motion information of the contiguous frames. Similar to the structure in [10, 17], the motion network is composed of multiple convolution layers and pooling layers, which are used to learn the feature representation of smoke images. In detail, it contains six convolution layers (marked as conv1, conv1_1, conv2, conv2_2, conv3, conv3_3, respectively) with the simplest stride of 2 in each of them and a tanh non-linearity after each layer. We set these parameters empirically. Taking the two concatenated consecutive two frames as input with a size of $h \times w \times 6$ (h is the frame height and w is the frame width). In this structure, the image resolution reduce to half of the original after each step, we repeat this operation three times so that the final size of the map is $\frac{1}{8}$ of the original one. There is a fully connected layer after the last pooling layer. The training details are described in the following sections.

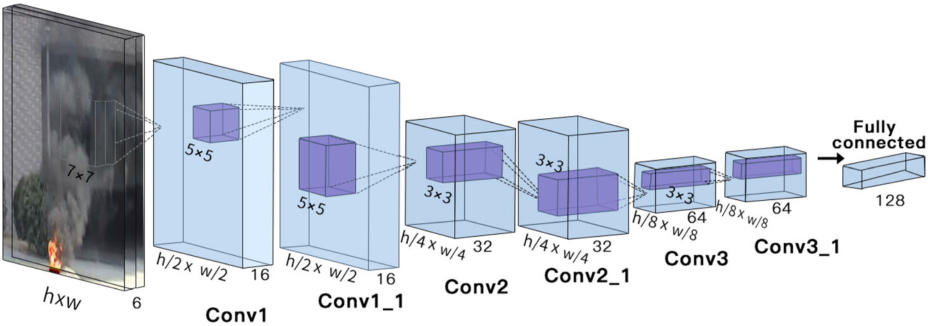


Fig. 6 The motion networks of our proposed RMSN consists of 6 convolutional layers (corresponding to Conv1, Conv1_1, Conv2, Conv2_1, Conv3 and Conv3_1) with stride of 2 in six of them and a tanh non-linearity layer after each layer. The inputs are two successive smoke frames with size of $h \times w \times 6$. And there is a fully-connected layer after the last convolutional layer. The purple cubes represent the convolutional kernels

3.2.2 Space network

Details of the space network are shown in Fig. 7, this network is used to learn the spatial characteristics of the original video frames. The network consists of three convolution layers and three pooling layers with a non-linearity tanh layer after each convolution layer. There is a fully connected layer after the last pooling one. The purple cubes are convolutional kernels and the red one is the pooling kernels. The stride in all convolution layers and pooling layers are set to 2. The original video RGB frames are the inputs of the network.

3.3 Motion-space feature fusion and aggregation of motion information

3.3.1 Motion-space feature fusion

There are different methods to fuse the two networks. Our aim is to fuse the motion-space characteristics in order to better obtain the motion-space characteristics of continuous video frames and the joint information between them. In order to best detect smoke, the shape and area of smoke will change with the influence of airflow and temperature, such as the upward motion direction, the gradually expanded area of smoke in the process of burning. We can argue that smoke is moving constantly and the movement is different from the ordinary fixed shape objects because the shape and size of these objects are the same in a continuous time series. Then, our motion network identifies the motion information in a continuous sequences and captures motion information from consecutive video frames. In this way, smoke can be identified by the combination of the two networks.

The fusion can be easily obtained when the feature maps of the two networks are of the same resolution at the layers to be fused. We choose to stack one network on another one. Suppose $X^A$and X^B are two feature maps from the motion and space networks layers that are needed to be fused, where W is the width, H is the height and D is the channel number of this two feature maps respectively. Y denotes the fused feature map. When applied to motion and space networks which consist of convolution, non-linearity and fully-connected layers, the fusion can be applied at different points in the two networks and it is easy to implement especially when the map dimensions are the same.

- Concatenation fusion: this fusion method stacks the feature maps of the same location i, j across the feature channels d together:

$$y^{cat}_{i,j,2d} = x^{A}_{i,j,d}, y^{cat}_{i,j,2d-1} = x^{B}_{i,j,d}, \tag{1}$$

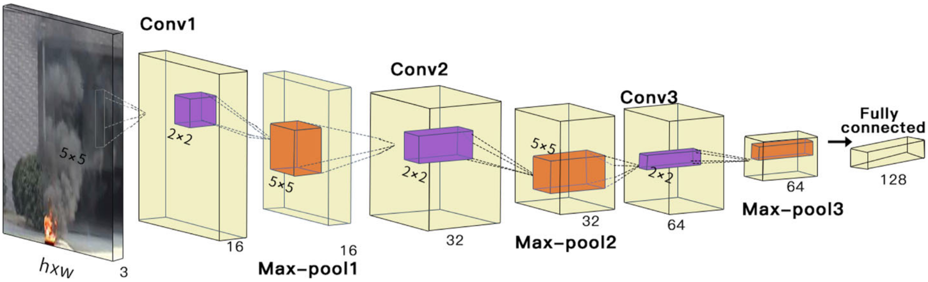


Fig. 7 Spatial networks of our proposed RMSN. It has 3 convolutional layers and 3 max-pooling layers with a tanh non-linearity layer interpolated after each convolutional layer. The fully-connected layer is the last layer of the network. The purple cubes are convolutional kernels while the red ones are pooling kernels

where $x^A, x^B \in R^{H\times W\times D}$, $y^{cat} \in R^{H\times W\times 2D}$ and $1 \leq i \leq H, 1 \leq j \leq W$.

- Sum fusion: this sum fusion method adds the feature maps of the same location i, j and channels d:

$$y^{sum}_{i,j,d} = x^A_{i,j,d} + x^B_{i,j,d}, \tag{2}$$

where $y^{cat} \in R^{H\times W\times 2D}$.

- Max fusion: this fusion method takes the maximum of the two feature maps:

$$y^{max}_{i,j,d} = max\{x^A_{i,j,d}, x^B_{i,j,d}\}. \tag{3}$$

In our experiments, we choose to perform the concatenation fusion operation after the fully connected layer because of the complexity of the two network structures. We compare the performance of the three fusion methods in terms of smoke detection accuracy.

3.3.2 Motion context information aggregation

We now consider how to combine the characteristics of the fused features which contains both spatial features and motion context information over time t. The motion information for each pair of smoke images is different because the length of the sequence is arbitrary. Therefore, we adopt the RNNs network which can process an arbitrary length of time sequence so that the problem of aggregating motion context information can be addressed by this neural network. In particular, the RNNs network has a feedback connection, which allows it to save information for a period of time and produce an output based on the information both of the current frame and the previous ones. The lateral connections serve as memory units, which allows the flow of information at any time step.

In the problem of smoke detection, the cumulative motion context information for the smoke detection is of great help because we need to learn abundant information of smoke images with different colors, shapes and densities. And the motion context information can be achieved by using the recurrent connections in the way that information can be passed over a long period. In other words, we aim to better extract the spatial and motion characteristics and the motion context information along video sequences, and then put these representative together to train a model to detect smoke. Given the output $f^{(t)} \in R^{p\times 1}$ of the fused spatial and temporal network p-dimension, the RNNs can be defined as follows:

$$o^{(t)} = Mf^{(t)} + Nr^{(t-1)}, \tag{4}$$

$$r^{(t)} = \tanh(o^{(t)}). \tag{5}$$

where $o^{(t)} \in R^{q\times 1}$ is the q-dimensional output of RNNs at time-step t, and $r^{(t-1)} \in R^{q\times 1}$ contains the information of all previous time steps. M $\in R^{q\times p}$ and N $\in R^{q\times p}$ represent the corresponding parameters for $f^{(t)}$ and $r^{(t-1)}$ respectively, where q is the dimension of the output of the last fully-connected layer in fusion part and p is the dimension of the feature embedding-space.

Finally, we added a temporal pooling layer after the RNNs, so that we can capture all the time information over the whole video clips. Temporal pooling is to obtain long-term information throughout the whole sequence, which combines motion context information captured through the RNNs. In this paper, we adopt mean-pooling over the temporal dimension to produce a single feature vector u to represent the spatial and motion information averaged over the whole raw sequence. The pooling method is as follows:

$$u = \frac{1}{T}\sum_{t=1}^{T} o^{(t)}, \tag{6}$$

where T is the length of the sequence or time-steps.

3.3.3 Loss function

In our smoke detection model, the loss function layer is the end with the input of two parts: the predicted value and the real label. The loss layer performs a series of operations on the two inputs to obtain the loss function $L(\theta)$ of the current network, where θ represents the vector of the current network weights. Our aim is to get a θ corresponding to the minimum $L(\theta)$. In this paper, we use the stochastic gradient descent method to optimize the approximation weight vector θ. The definition of L is as follows:

$$L(\theta) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{2} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{2} e^{\theta_l^T x^{(i)}}}\right] \tag{7}$$

Where N is the batch size.

4 Experiments

4.1 Experimental settings

(1)Datasets We test the proposed model on three widely used public video datasets from B.C. Ko [14], Toreyin et al. [32], R. Vezzani [33] and our large-scale dataset. The details of the four datasets are described in Table 1.

(2)Baselines To show the superiority of the proposed RMSN, we compare our method with other three methods ones based on AlexNet [15] model, texture features, and Haar-like features [38]. To evaluate the effectiveness of our model in combing the concatenation fusion methods, three fusion algorithms are used for comparison. Finally, RNNs can make full use of the cumulative context information of the whole sequence. To show the advantages of the recurrent units, we implement the proposed model with and without recurrent units.

(3)Implementation details All the input pairs of video frames are resized to 128 × 64 and we choose Adam [13] as the optimization method due to its faster convergence than standard stochastic gradient descent. The iteration times is set to 10000 to optimize the model and the training model is saved once every 1000 times. Our network is trained by the BP algorithm with batch stochastic gradient descend and the soft-max loss is minimized. The weight decay is set to 0.0005 and learning rate is 0.01 in all models. All weights are initialized from a zero-centered normal distribution with standard deviation 0.01. We separate each dataset into a training set and a testing set, with about 2:1 ratio to train and test all models. All

Table 1 The details of dataset used in the paper

Dataset	Number of videos	Average length/(s)	Description
B.C.Ko [14]	16	120	6 smoke videos, 10 smoke-like videos
Toreyin et al. [32]	20	147	Indoor and outdoor smoke videos
R. Vezzani [33]	14	42	Outdoor smoke videos
Ours	5000	80	3000 smoke and 2000 non-smoke videos

Table 2 Comparison results of the proposed method and three other methods in the B.C. Ko [14] dataset in terms of TPR and TNR

B.C.Ko [14]	Ours	AlexNet [15]	Haar-like [38]	Texture
TPR	0.9525	0.9220	0.9155	0.8865
TNR	0.9853	0.9432	0.9325	0.9028

parameters are set empirically. Experimental results show the highest classification accuracy with these parameters. The processing time for a frame is less than 30 ms, in other words, our method can process videos with size of 128 × 64 at above 33 frames per seconds.

(4)Evaluation protocol All the algorithms are evaluated based on the following widely-used criteria. Firstly, we use TPR and TNR to evaluate our method, (8) and (9) show how we get them:

$$TPR = \frac{TP}{(TP + FN)} \tag{8}$$

$$TNR = \frac{TN}{(TN + FP)} \tag{9}$$

TP (true positive) is the number of correct detections of smoke, FN (false negative) is the number of frames which have smoke but not recognized. FP (false positive) is the number of frames which do not have smoke but recognized as smoke, TN (true negative) is the number of correct detections of non-smoke.

In addition, to evaluate the performance of smoke detection, we exploit the receiver operator characteristic (ROC). The ROC curves are obtained by experimenting in our dataset. For each experiment, we plot the true positives rate (TPR) on the Y-axis and plot the false positive rate (FPR) value on the X-axis. The nearer the ROC curve is to the upper left corner, the higher the accuracy of the test is.

4.2 Experimental results and analysis

4.2.1 Comparison with the state-of-art methods

We compare the performance of our method based on RMSN model with other three ones based on AlexNet [15] model, texture features, and Haar-like features [38]. In Table 2, we show the comparison results of our method with the other three methods as described before on the whole Keimyung dataset. Additionally, we also compare our results with the same methods on Bilkent dataset, the comparison results are shown in Table 3. Furthermore, experimental results of Modena dataset are described in Table 4.

As shown in Tables 2, 3 and 4, our method outperforms all other smoke detection methods. Compared with the second best approach based on AlexNex [15] model, our method has slightly higher precision both in TPR and TNR. The AlexNet [15] model can learn effective representation of smoke images but without motion information. In contrast, our method can learn discriminative spatial representation and motion context information of

Table 3 Experimental results of four different methods in the Toreyin et al. [32] dataset in terms of TPR and TNR

Toreyin et al. [32]	Ours	AlexNet [15]	Haar-like [38]	Texture
TPR	0.9575	0.9280	0.9175	0.8910
TNR	0.9895	0.9485	0.9265	0.9055

Table 4 TPR and TNR results in the R. Vezzani [33] dataset

R. Vezzani [33]	Ours	AlexNet [15]	Haar-like [38]	Texture
TPR	0.9530	0.9260	0.9178	0.8892
TNR	0.9780	0.9555	0.9245	0.9025

the whole sequence which performs better even when the datasets are limited. This shows the robustness of the proposed method since false alarms are considerably reduced.

Simultaneously, we train our own model in our new dataset which consists of large smoke and small smoke. Our experiments are as follows. One group is for large smoke, another group is for small smoke. Lastly, we use all of the smoke samples. The comparison of the four methods on our dataset is shown in the Figs. 8 and 9.

From Figs. 8 and 9, we can conclude that both the TPR and TNR of our method are higher than other three methods. In the experiment of all smoke samples, our method outperforms other methods with an average TPR of 94.85%. Besides, we obtain an average TNR of 97.20%. We can conclude that the smoke detection accuracy is relative low of the traditional method based on texture features because this method is vulnerable to light, airflow and other factors. Although Haar-like [38] method extract haar-like features from integral images combined with dynamic analysis to reduce false alarm rate, it can not perform well in complex and intermittent frames. The AlexNet [15] model can learn effective features of smoke images but without motion information. Obviously, TPR and TNR of other three methods are relatively low in our dataset because these methods do not have good generalization, especially in complex scenes. In contrast, our method can learn both spatial features and motion context information of the whole sequence which performs better in complex scenes. This shows higher generalization of the proposed method because of the high TPR and TNR in diverse scenes. It is important to note that these promising results are achieved by exploiting both spatial and motion context information over time. This indicates that a large number of representative motion-space features are extracted to improve the accuracy of smoke detection.

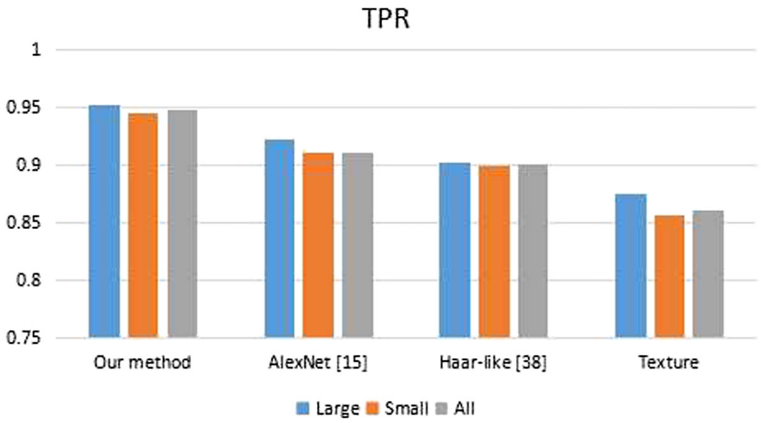


Fig. 8 Positive samples. Experimental results of different smoke detection methods in terms of TPR of positive testing samples

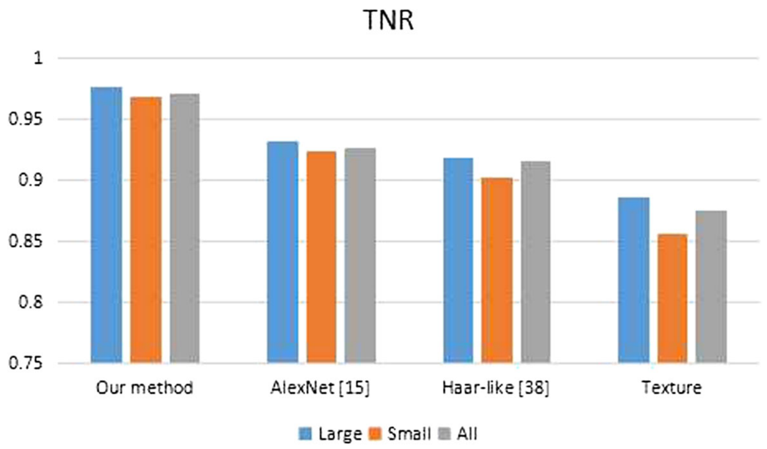


Fig. 9 Negative samples. TNR of non-smoke testing samples

Furthermore, performances of different methods on our large smoke videos, small smoke videos and all smoke videos are evaluated by ROC curves. The ROC curves of the four different methods are shown in Figs. 10, 11 and 12. It is obvious that the ROC curves of the method based on texture features locate in the lowest position of each experiment. The ROC curves of our method are nearest to upper left corner of the chart among all curves. This notes that our method performs better in smoke detection.

Additionally, for outdoor smoke detection in complex sceneries, video-based smoke detection performs better than traditional photo-based method. The proposed method is compared with those of the same kind video based methods described before. In Fig. 13,

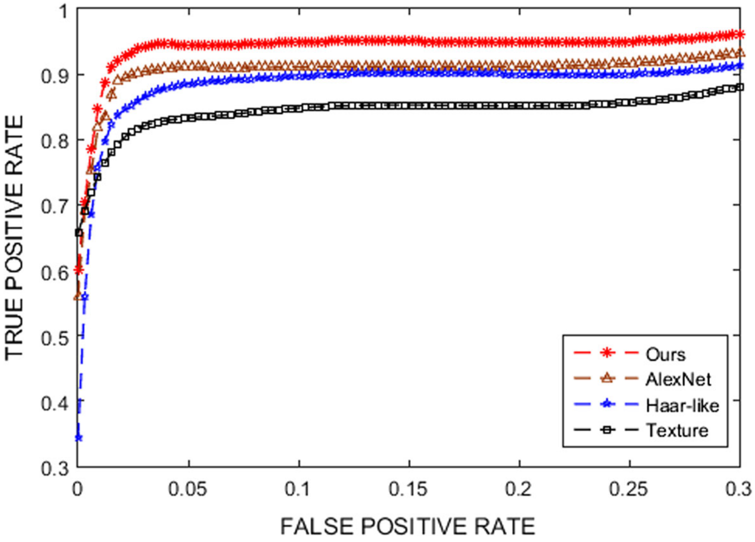


Fig. 10 ROC curves on large smoke video clips. To show the advantage of RMSN, we apply the proposed method RMSN and other three methods on our large smoke videos

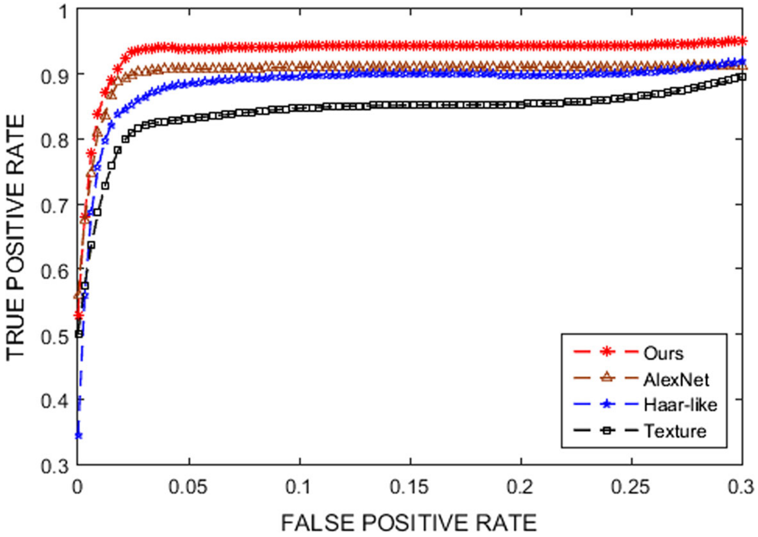


Fig. 11 The ROC curves of the proposed RMSN method and other three state-of-art methods on our small smoke video clips

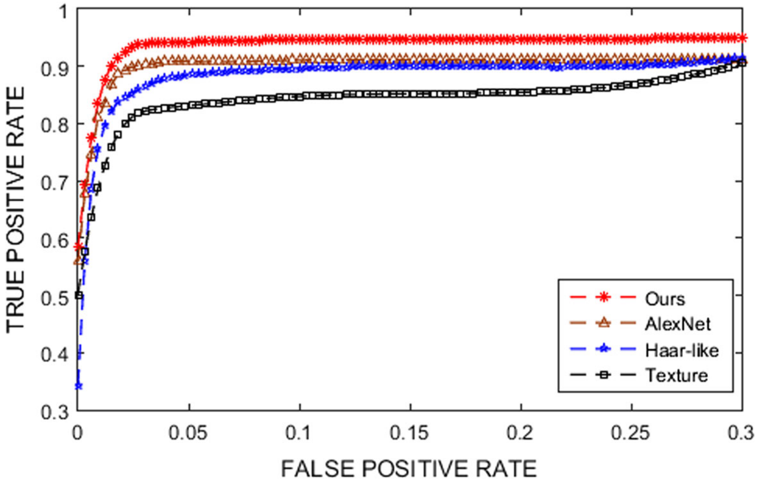


Fig. 12 Curves of ROC for the proposed method and other three different methods on all our dataset

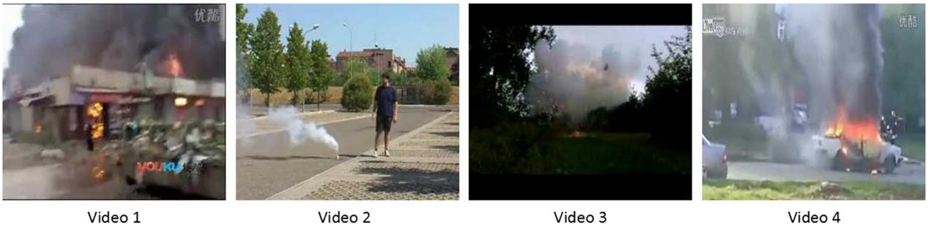


Fig. 13 Testing videos: Video1 factory smoke, Video2 road smoke, Video3 forest smoke and Video4 car smoke

Table 5 Smoke detection performance comparisons on videos

Smoke videos	Duration	Detection frame numbers			
		Ours	AlexNet [15]	Haar-like [38]	Texture
Video1 factory smoke	1789	1	2	2	10
Video2 road smoke	1031	384	401	415	480
Video3 forest smoke	210	25	34	35	51
Video4 car smoke	524	19	22	26	41

Video1,Video2,Video3 and Video4 illustrates a factory smoke, road smoke, forest smoke and car smoke. As shown in Table 5, our method has a better performance than the other three methods when the smoke videos are tested. After observing alarm frame numbers, we can conclude that our method can provide an earlier smoke alarm. The reason that our method can achieve a better performance is mainly the robustness of both space and motion context features. But the processing speed of our method is slower than AlexNet [15].

At the same time, we present some instances that are falsely classified by our model training by all smoke images, as shown in Fig. 14. Smoke instances are falsely classified as non-smoke ones in Fig(a). In another way, there are many non-smoke images similar to smoke in appearance. Fig(b) shows the non-smoke ones that are classified as smoke. The samples shown in Fig(b) are particularly similar to smoke in shapes, color and texture. It is of great difficulty to distinguish smoke and non-smoke images accurately because it is even hard for people to identify in some extent.

4.2.2 Advantages of motion networks

As described in Section 3, a pair of consecutive video frames of smoke is processed by the motion network within RMSN (corresponding to the blue rectangles in Fig. 5) to predict

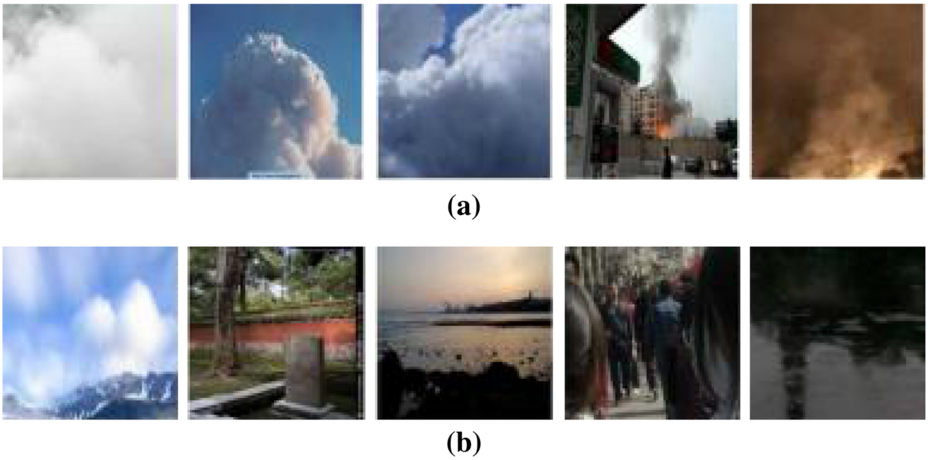


Fig. 14 Falsely classified smoke and non-smoke videos. **a** illustrates smoke instances falsely classified as non-smoke ones. Non-smoke samples with false alarms which are similar to smoke in appearance are shown in (**b**)

Table 6 Comparison between different methods with and without motion networks

Methods	B.C.Ko [14]		Toreyin et al. [32]		R. Vezzani [33]		BJTU-Smoke	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Ours	0.9525	0.9853	0.9575	0.9895	0.9530	0.9780	0.9485	0.9720
Our method without motion net	0.9422	0.9725	0.9445	0.9785	0.9365	0.9635	0.9355	0.9625

motion between the adjacent frames at each time-step. We will compare our method with and without motion networks in the above dataset. The results are shown in Table 6.

We can notice that our model outperforms the one without motion networks in all dataset. The performance is boosted by 1.03%, 1.3%, 1.65% and 1.3% of TPR and 1.28%, 1.10%, 1.45% and 0.95% of TNR in the above dataset respectively. Although RNNs units can aggregate the motion context information, the motion networks can predict motion between the adjacent frames at each time-step by processing a pair of consecutive video frames and output abundant motion context information to RNNs.

4.2.3 Comparison between different fusion methods

In this part, we compare three different fusion methods in Table 7. From the first three rows of Table 7, we demonstrate the smoke detection rate TPR and TNR of the model trained on small smoke, large smoke and all smoke images of our dataset. From the results, we know that the concatenation method adopted in this paper performs better than the other two methods in some degree. The reason may be that the concatenation fusion can better maintain the original features.

4.2.4 Advantages of recurrent units

We also compare our method with and without recurrent units in the last row of Table 7. In the traditional neural network, it is assumed that all inputs or outputs are independent of each other. In contrast, RNNs can make full use of the information of the whole sequence which make a decision according to both the cumulative context information and the current input.

From the first and the last row of Table 6, we can conclude that our proposed model performs better than the model without recurrent units. When these recurrent units are applied in our model, the performance is boosted by 2.30%, 1.95% and 2.00% of TPR and 2.60%,

Table 7 Comparison between different fusion methods and the model without recurrent units

Methods	Small smoke		Large smoke		All	
	TPR	TNR	TPR	TNR	TPR	TNR
Concatenation fusion	0.9450	0.9685	0.9520	0.9765	0.9485	0.9720
Sum fusion	0.9250	0.9420	0.9335	0.9450	0.9275	0.9435
Max fusion	0.9325	0.9475	0.9385	0.9490	0.9355	0.9485
Concatenation fusion without recurrent units	0.9220	0.9425	0.9325	0.9475	0.9285	0.9455

2.90% and 2.65% of TNR in small, large and all smoke images respectively. By aggregating the motion context information, our method can better distinguish whether the moving objects are smoke, because the motion direction of the smoke is ascending by the influence of the heat and the shapes and areas of the smoke are affected by airflow. It is obvious that the cumulative motion context information of the whole sequence captured by the recurrent units is of great help for smoke detection.

5 Conclusion

It is difficult to find the appearance of smoke at the early stages of smoke because the area is relatively small and it is vulnerable to interference. Smoke detection method based on feature extraction can not extract the features of smoke effectively resulting in high false alarm rate and low detection rate.

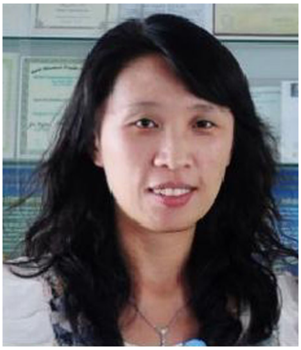
To improve the robustness and adaptability of smoke detection, we create a large-scale smoke dataset, going beyond previous ones. Moreover, a novel smoke detection model learning spatial representation and motion context information from source video frames is proposed. Experimental results, carried out under various challenging conditions, demonstrate that our recurrent motion-space context model is beneficial for the smoke detection accuracy. In the future work, we will continue to study how to improve smoke detection by incorporating additional discriminant information.

References

1. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2016) Apparent age estimation from face images combining general and children-specialized deep learning models. In: IEEE conference on computer vision and pattern recognition workshops, pp 801–809
2. Avgerinakis K, Briassouli A, Kompatsiaris I (2012) Smoke detection using temporal hoghof descriptors and energy colour statistics from video. In: International workshop on multi-sensor systems and networks for fire detection and management
3. Bao BK, Liu G, Xu C, Yan S (2012) Inductive robust principal component analysis. IEEE Trans Image Process 21(8):3794–3800
4. Bao BK, Zhu G, Shen J, Yan S (2013) Robust image analysis with sparse representation on quantized visual features. IEEE Trans Image Process 22(3):860–871
5. Barmpoutis P, Dimitropoulos K, Grammalidis N (2014) Smoke detection using spatio-temporal analysis, motion modeling and dynamic texture recognition. In: Signal processing conference, pp 1078–1082
6. Chen J, Wang Y, Tian Y, Huang T (2013) Wavelet based smoke detection method with rgb contrast-image and shape constrain. In: Visual communications and image processing, pp 1–6
7. Chen J, You Y, Peng Q (2013) Dynamic analysis for video based smoke detection. International Journal of Computer Science Issues
8. Chen TH, Yin YH, Huang SF, Ye YT (2006) The smoke detection for early fire-alarming system base on video processing. In: International conference on intelligent information hiding and multimedia signal processing, pp 427–430
9. Cui Y, Dong H, Zhou E (2008) An early fire detection method based on smoke texture analysis and discrimination. In: Congress on image and signal processing, 2008. CISP '08, pp 95–99
10. Dosovitskiy A, Fischer P, Ilg E, Husser P, Hazirbas C, Golkov V, Smagt PVD, Cremers D, Brox T (2015) Flownet: learning optical flow with convolutional networks. In: IEEE international conference on computer vision, pp 2758–2766
11. Genovese A, Labati RD, Piuri V, Scotti F (2011) Wildfire smoke detection using computational intelligence techniques. In: IEEE international conference on computational intelligence for measurement systems and applications, pp 1–6
12. Iida Y, Maruta H, Kurokawa F (2013) A study on smoke detection method based on lbp featured and adaboost. Ieice Technical Report Image Engineering 112(475):57–62

13. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. Computer Science
14. Ko BC (2012) Wildfire smoke detection using temporospatial features and random forest classifiers. Opt Eng 51(1):7208
15. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: International conference on neural information processing systems, pp 1097–1105
16. Krstini D, Stipaničev D, Jakovčevi T (2015) Histogram-based smoke segmentation in forest fire detection system. Information Technology and Control 38(3):237–244
17. Liu H, Jie Z, Jayashree K, Qi M, Jiang J, Yan S, Feng J (2017) Video-based person re-identification with accumulative motion context. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
18. Liu S, Ou X, Qian R, Wang W, Cao X (2016) Makeup like a superstar: deep localized makeup transfer network. In: International joint conference on artificial intelligence, pp 2568–2575
19. Liu S, Sun Y, Zhu D, Bao R, Wang W, Shu X, Yan S (2017) Face aging with contextual generative adversarial nets. In: ACM, pp 82–90
20. Liu S, Wang C, Qian R, Yu H, Bao R (2017) Surveillance video parsing with single frame supervision. In: IEEE conference on computer vision and pattern recognition workshops, pp 1–9
21. Lu J, Wang G, Deng W, Moulin P (2015) Multi-manifold deep metric learning for image set classification. In: IEEE conference on computer vision and pattern recognition, pp 1137–1145
22. Maruta H, Nakamura A, Kurokawa F (2010) A new approach for smoke detection with texture analysis and support vector machine. In: IEEE international symposium on industrial electronics, pp 1550–1555
23. Millangarcia L, Sanchezperez G, Nakano M, Toscanomedina K, Perezmeana H, Rojascardenas L (2012) An early fire detection algorithm using ip cameras. Sensors 12(5):5670–86
24. Niu Z, Zhou M, Wang L, Gao X, Hua G (2016) Ordinal regression with multiple output cnn for age estimation. In: IEEE conference on computer vision and pattern recognition, pp 4920–4928
25. Park JO, Ko BC, Nam JY, Kwak SY (2013) Wildfire smoke detection using spatiotemporal bag-of-features of smoke. In: IEEE workshop on applications of computer vision, pp 200–205
26. Tian H, Li W, Ogunbona P, Wang L (2014) Single image smoke detection. In: Europeon conference on computer vision
27. Tian H, Li W, Wang L (2012) Ogunbona: a novel video-based smoke detection method using image separation. In: IEEE international conference on multimedia and expo, pp 532–537
28. Tian H, Li W, Wang L, Ogunbona P (2012) A novel video-based smoke detection method using image separation. In: IEEE international conference on multimedia and expo, pp 532–537
29. Tian H, Li W, Wang L, Ogunbona P (2014) Smoke detection in video: an image separation approach. Int J Comput Vis 106(2):192–209
30. Tian Y, Luo P, Wang X, Tang X (2015) Pedestrian detection aided by deep learning semantic tasks. In: IEEE conference on computer vision and pattern recognition, pp 5079–5087
31. Tian Y, Luo P, Wang X, Tang X (2015) Deep learning strong parts for pedestrian detection. In: IEEE international conference on computer vision, pp 1904–1912
32. Toreyin BU, Dedeoglu Y, Cetin AE (2005) Wavelet based real-time smoke detection in video. In: Signal processing conference, 2005 european, pp 1–4
33. Vezzani R, Calderara S, Piccinini P, Cucchiara R (2008) Smoke detection in video surveillance:the use of visor (video surveillance on-line repository). In: ACM international conference on image and video retrieval, Civr 2008, Niagara Falls, Canada, July, pp 289–298
34. Wu J, Yu Y, Huang C, Yu K (2015) Deep multiple instance learning for image classification and auto-annotation. In: IEEE conference on computer vision and pattern recognition, pp 3460–3469
35. Yang S, Zheng X (2014) A video smoke detection method based on various features integration and adaboost. J Comput Inf Syst 10(24):10,463–10,471
36. Yuan F (2008) A fast accumulative motion orientation model based on integral image for video smoke detection. Pattern Recogn Lett 29(7):925–932
37. Yuan F (2011) Video-based smoke detection with histogram sequence of lbp and lbpv pyramids. Fire Safety Journal 46(3):132–139
38. Yuan F, Fang Z, Wu S, Yang Y (2015) Real-time image smoke detection using staircase searching-based dual threshold adaboost and dynamic analysis. IET Image Process 9(10):849–856
39. Zhu Z, Liang D, Zhang S, Huang SX, Li B (2016) Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2110–2118
40. Zhao Y, Zhou Z, Xu M (2015) Forest fire smoke video detection using spatiotemporal and dynamic texture features. J Electr Comput Eng 2015(3):40

Mengxia Yin graduated in School Of Computer and Communication Engineering from Zhengzhou University of Light Industry, Henan, China, in 2014. She is a Master Degree Candidate with School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her research interests include image processing and computer vision.



Congyan Lang received her Ph.D. from Beijing Jiaotong University in 2006. She is currently a Full Professor in the School of Computer and Information Technology, Beijing Jiaotong University. She has been a Visiting Scholar in the Department of Electrical and Computer Engineering, National University of Singapore (2010–2011) and the Department of Computer Science, University of Rochester(2013–2014), respectively. Her research interests include machine learning and computer vision.

Zun Li is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University (BJTU), Beijing, China. Her research interests focus on the image processing, computer vision, and machine learning etc.

Songhe Feng received the Ph.D. Degree in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, PR China, in 2009. He is currently an Associate Professor in the School of Computer and Information Technology, Beijing Jiaotong University. He has been a visiting scholar in the Department of Computer Science and Engineering, Michigan State University (2013–2014). His research interests include computer vision and machine learning.

Tao Wang received the Ph.D. Degree in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, P.R. China, in 2013. He is currently an Associate Professor in the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include graph algorithm, pattern recognition and image understanding.