Multimed Tools Appl
https://doi.org/10.1007/s11042-018-5978-5

Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features

Yaocong Hu[1,2] · **Xiaobo Lu**[1,2]

Received: 16 November 2017 / Revised: 29 March 2018 / Accepted: 5 April 2018

Abstract Fire is one of the most dangerous disasters threatening human life and property globally. In order to reduce fire losses, researches on video analysis for early smoke detection have become particularly significant. However, it is still a challenging task to extract stable features for smoke recognition, largely due to its variations in color, shapes and texture. Classical convolutional neural networks can automatically learn feature representations of appearance from a single frame but fail to capture motion information between frames. For addressing this issue, in this paper, we propose a spatial-temporal based convolutional neural network for video smoke detection, and for real-time detection, propose an enhanced architecture, which utilizes a multitask learning strategy to jointly recognize smoke and estimate optical flow, capturing intra-frame appearance features and inter-frame motion features simultaneously. The effectiveness and efficiency of our proposed method is validated by experiments carried out on our self-created dataset, which achieves 97.0% detection rate and 3.5% false alarm rate with processing time of 5ms per frame, obviously outperforming existing methods.

Keywords Smoke detection · Convolutional neural networks · Spatial-temporal · Multi-task learning

This work was supported by the National Key Science & Technology Pillar Program of China (No. 2014BAG01B03), the National Natural Science Foundation of China (No. 61374194), Key Research and Development Program of Jiangsu Province (No. BE2016739), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

✉ Xiaobo Lu
xblu2013@126.com

[1] School of Automation, Southeast University, Nanjing 210096, China

[2] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China

Published online: 14 April 2018

1 Introduction

Fire is one of the most frequently happened and extremely destructive disasters. The sooner the fire is detected, the better the chances are for survival. To forecast the potential fire hazards and effectively reduce the damages, smoke detection systems were developed to protect against the occurrence of fire accidents. Traditional smoke detection systems were generally sensors based which sampled temperature, humidity or carbon monoxide concentration in real time [2, 14]. Such sensors have been widely used due to their low cost and convenient use, but some inherent limitations of these sensors exist that are difficult to overcome. First, these sensors require to be installed closely to the fire and are easily to be damaged in high temperature. Additionally, they are limited to be applied in small or indoor spaces. In order to overcome the above mentioned limitations of traditional smoke detection system, a brand new smoke detection system based on surveillance video has become mainstream in the past few years, which has remarkable advantages over traditional sensor-based systems. Meanwhile, varieties of computer vision-based algorithms have been proposed for video smoke detection in the literature, which can recognize smoke automatically instead of human monitoring. Most of existing smoke detection algorithms are based on traditional handcrafted features, approaching from a number of angles, but sharing a common framework: feature extraction and classification. The flowchart of handcrafted feature-based algorithms is shown in Fig. 1. In training stage, they first extract color, shape and texture information from training images and combine these information into multi-dimensional features; then, use the extracted features as an input to learn a supervised classifier, such as Support Vector Machine (SVM), Adaboost and neural network, etc. For testing, they extract the same

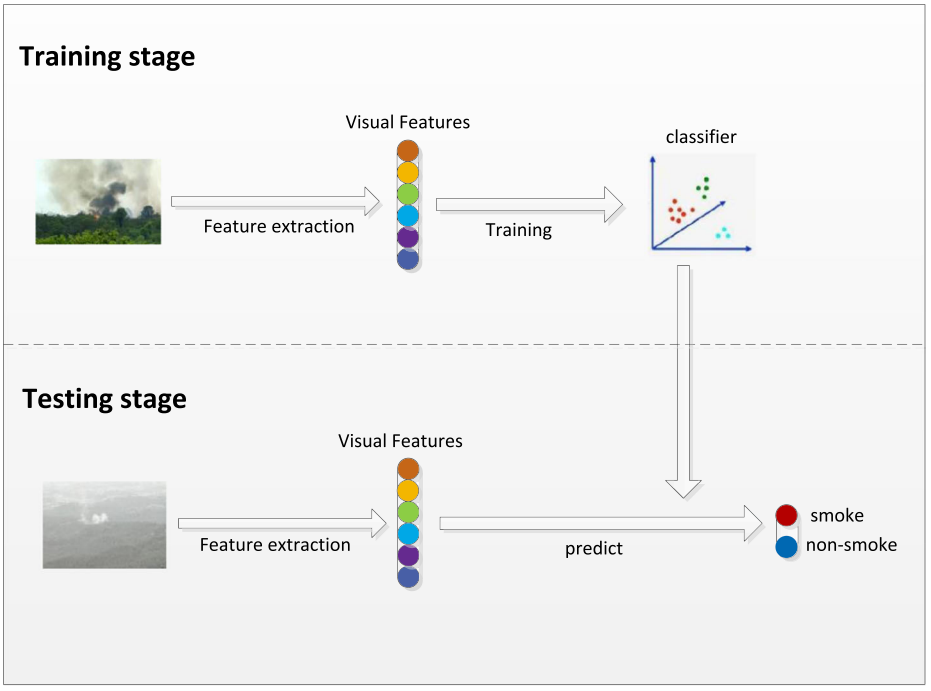

Fig. 1 The flowchart of traditional smoke detection algorithm

features as in training stage and the trained classifier judges whether an image contains fire smoke.

For traditional smoke detection algorithms as mentioned above, they have following obvious disadvantages: (1) the performance of these algorithms largely depends on features which are extracted and combined manually, but these features are selected mainly empirical and can not be applicable in all scenes due to huge variations of smoke in color, shapes and texture. (2) it is extremely difficult to achieve high detection rate and low false alarm rate simultaneously; the main reason is that these algorithms cannot correctly distinguish smoke with other objects which share similar characteristics to smoke. As illustrated in Fig. 2, the top row shows the smoke images, and the bottom row shows the non-smoke images. However, clouds in Fig. 2c and fogs in d are mistakenly recognized to smoke by using the method of [28].

Compared with the handcrafted feature-based algorithms, Convolutional Neural Network (ConvNet) structure is usually composed of multi-layers and can automatically learn a unique set of features for a given task. In recent years, ConvNet architecture has achieved outstanding performance in many visual processing tasks, including image and video classification [15, 17, 36], object and face detection [5, 25], crowd analysis [10, 37], speech recogonition[22], etc. The remarkable achievements of ConvNet on visual processing are largely contributed to its excellent abstract ability. Therefore, it is reasonable to apply the power of ConvNet features to the specific task of smoke detection instead of using traditional handcrafted features.

Inspiring by the great success of ConvNet in multiple computer vision tasks, Tao et al. [26] first applied AlexNet [17] to video smoke detection and achieved the state-of-the-art performance. AlexNet which made a breakthrogh success in Large Scale Visual Recognition Challenge(LSVRC) [21] can automatically learn feature representations of appearance from a single frame but is unable to capture motion information between frames. Thus, there is definitely rooms for improvements, as long as we design a deep architecture which can

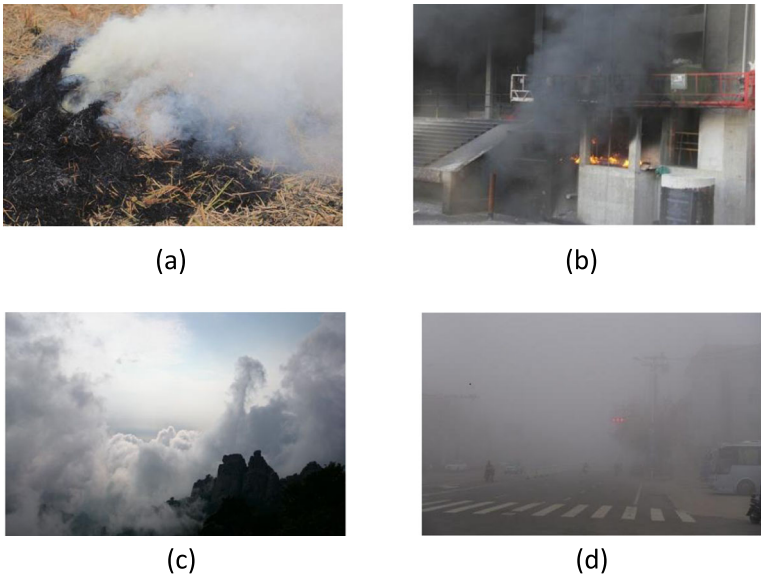

Fig. 2 Examples of positive and negative samples

extract complementary motion features. To realize this demand, in this paper, we train a two stream (spatial-temporal stream) convolutional neural network extracting appearance and motion features simultaneously, and then propose an enhanced architecture for real time detection. The main contributions of our studies can be concluded into three aspects:

- For video smoke detection, we train a two stream (spatial-temporal stream) convolutional neural network which can extract both intra-frame appearance and inter-frame motion information from video sequence.
- For real time video smoke detection, we propose an enhanced spatial-temporal architecture, which utilizes a multitask learning method to recognize smoke and estimate optical flow simultaneously.
- We test our proposed method on our self-created dataset which achieves the state-of-the-art result.

The rest of this paper is organized as follows: we review the previous works about smoke detection in Section 2; the proposed methods and the overall frameworks are detailed in Section 3; experiments and the comparisons of results are summarized in Section 4; finally, we conclude this paper in Section 5.

2 Related work

Specific object recognition and detection has long been a hot topic in computer vision researches [7, 8, 11, 30]. In recent years, several solutions for video smoke detection have been proposed. According to the features used in their analysis, we can roughly divide existing methods into two categories of handcrafted feature-based methods [6, 16, 24, 27, 28, 32–35] and deep learning based methods [4, 26, 29, 31, 38]. In this section, we briefly review the corresponding methods.

Handcrafted feature-based methods of smoke detection Toreyin et al. [28] presented a video smoke detection method by using combined features of color, edge, motion and wavelet. Srisuwan et al. [24] proposed a framework which employed local Gray Level Co-occurrence Matrix (GLCM) features and global motion features, then a BP neural network were used to classify candidate smoke regions . Tian et al. [27] extracted Non-Redundant Local Binary Pattern(NRLBP) features from consecutive frames and the feature vectors of positive and negative samples were used to train a Support Vector Machine(SVM). Ko et al. [16] utilized a spatial-temporal Bag of Words (BoW) models to construct visual words histogram in each local region, then smoke or non-smoke frames were judged by random forests. Yuan et al. in [32] proposed a solution based on accumulative motion orientation model and an integral image calculation accelerated smoke detection to real-time, in [34] proposed a double mapping framework combined with Adaboost for smoke detection.

In general, we can summarize these methods as below: manually extracting combined features of color, motion and texture, followed by a classifier to judge smoke or non-smoke. However, smoke has so various in color, shapes and other characteristics that difficult to be well-represented by handcrafted features.

Deep learning based methods of smoke detection Tao et al. [26] trained an end-to-end Alexnet to extract features and map the raw images to classifier outputs automatically. Frizzi et al. [4] presented a solution based on convolutional neural networks and smoke region can be localized by converting fully connected layers to convolutions. Yin et al. [31]

took advantages of batch normalization strategy to speed up convergence in training stage and improve performance in test stage. Zhang et al. [38] proposed a cascade ConvNet model in which the global image-based Convnet was used for classification and the local patch-based ConvNet was used for localization. Xu et al. [29] utilized synthesis smoke to extend training samples and then proposed a deep domain adaption method for ConvNet learning.

To sum up the above mentioned methods, they all trained end-to-end deep learning architectures which can automatically learn appearance feature representations from single frames without manually designing features and they achieved obvious improvements on performance criterions compared with handcrafted feature-based methods. But they are suboptimal due to failing to capture motion features between frames.

Spatial-temporal stream ConvNet Simonyan et al. [23] first proposed a two stream-(spatial-temporal) ConvNet model for action recognition in video clips. In their implementations, spatial stream extracts appearance features from a single frame, while temporal stream extracts motion features from stacked optical flow between multi consecutive frames. The two streams are trained separately and then combined by late SVM fusion.

For video smoke detection, motion features convey the speed and orientation of smoke diffusion which are well complementary to static appearance features. Thus the combination of both appearance and motion features may bring to a significant boost in our performance. But it should be noted that we can not stack optical flow between multi frames as in [23], because in our application of surveillance video, smoke or non-smoke need to be recognized frame by frame.

To the best of our knowledge, so far, this is the first publication applying spatial-temporal ConvNet architecture to the video smoke detection task. Additionally, for real time detection, we enhance the spatial-temporal architecture which utilize a multitask strategy to jointly learn appearance and motion features. The effectiveness and efficiency of our proposed method will be proven by experiments in Section 4 and the specific implementations and formulations are introduced in next Section 3.

3 Methodology

In this section, we give a detailed description of our proposed architectures which can capture appearance information together with motion information, aiming at determine whether there is smoke or not. On the basis of this idea, we give two available proposals: (1) utilize the two stream(spatial-temporal stream) ConvNet for video smoke detection where two streams are trained separately and then combined by late SVM fusion; (2) utilize an enhanced spatial-temporal architecture adopting multitask learning strategy to jointly recognize smoke and estimate optical flow simultaneously in a single stream. We respectively introduce the details of the two proposed architecture in next Sections 3.1 and 3.2.

3.1 Proposal 1: spatial-temporal stream ConvNet for video smoke detection

We can roughly decompose video sequence into two components; in spatial component, raw images convey appearance information of objects such as color, shapes or texture; while, in temporal component, optical flow between neighbor frames is utilized to represent motion information of objects in a scene. Naturally, spatial component and temporal component is discriminatively fed into spatial stream and temporal stream for training. For testing, each stream separately performs feed forward and generates classification scores which are then

combined by a late SVM fusion. An illustration of the spatial-temporal stream Convnet used for video smoke detection is shown in Fig. 3 and we list the parameters of each layer in Table 1.

Spatial stream Spatial stream ConvNet effectively classifies smoke or non-smoke from still frames and takes similar structure to that in [17] which composed of eight layers. The first five layers are convolutional layers and the last three layers are fully connected layers(inner product layers). Relu, pooling and normalization follows after the output of each convolutional layer. We take a square RGB image of $227\times227\times3$ as input and the first convolutional layer filters input with 96 kernels of $11\times11\times3$. The second convolutional layer takes as input the output of the first convolutional layer and filters it with 256 kernels of size $5\times5\times96$. The third convolutional layer has 384 kernels of size $3\times3\times256$. Note that the parameters of other layers can be looked up in Table 1. Unlike the Alexnet, the last full connected layer has 2 neurons which outputs the probability of smoke and non-smoke.

Temporal stream In contrast with spatial stream, temporal stream ConvNet utilizes optical flow displacement fields between neighbor two frames as input. Such input provides direction and magnitude of each pixel motion, which is a good complementary to static appearance information. For consecutive two frames, we use the Brox et al.'s method [1] to compute dense optical flow and decompose the motion into horizontal and vertical components as we can see in Fig. 4. Figure 4a and b shows the consecutive two frames. Figure 4c

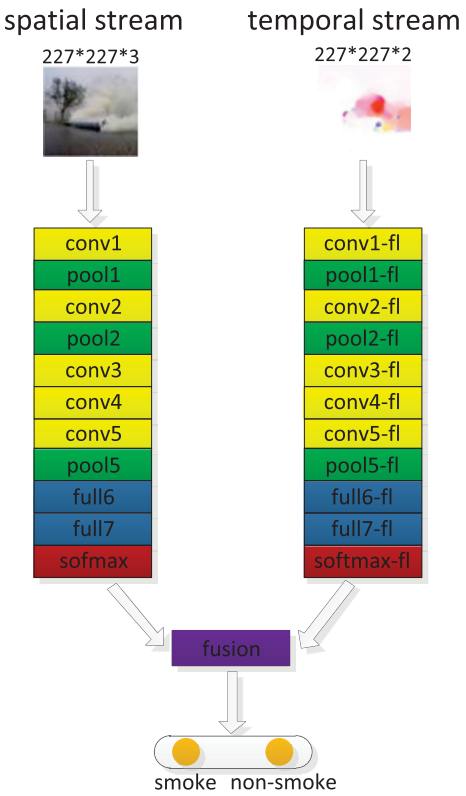

Fig. 3 Architecture of proposal 1: spatial-temporal stream ConvNet

Table 1 The relevant layers and their parameters of proposals 1

Layer	Layer Type	Parameter
input	input	Image size: 227×227×3
conv1	convolution	Filter size: 11×11, Filter number: 96, Stride:4
pool1	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv2	convolution	Filter size: 5×5, Filter number: 256, Stride:1
pool2	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv3	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv4	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv5	convolution	Filter size: 3×3, Filter number: 256, Stride:1
pool5	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
full6	fully-connected	Neurons output: 4096
full7	fully-connected	Neurons output: 2048
softmax	softmax	Neurons output: 2
input-fl	input	Image size: 227×227×2
conv1-fl	convolution	Filter size: 11×11, Filter number: 96, Stride:4
pool1-fl	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv2-fl	convolution	Filter size: 5×5, Filter number: 256, Stride:1
pool2-fl	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv3-fl	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv4-fl	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv5-fl	convolution	Filter size: 3×3, Filter number: 256, Stride:1
pool5-fl	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
full6-fl	fully-connected	Neurons output: 4096
full7-fl	fully-connected	Neurons output: 2048
softmax-fl	softmax	Neurons output: 2

is the horizontal component of the displacement vector filed and Fig. 4d is the vertical component of the displacement vector filed. Figure 4e shows the vector field of combined horizontal and vertical components. Formally, we define the horizontal component as d^x and the vertical component as d^y, and the two input channels of temporal stream I_{fl} can be represented as follow:

$$\begin{aligned} I_{fl}(u, v, 1) &= d^x(u, v), \\ I_{fl}(u, v, 2) &= d^y(u, v), u = [1, 2..., w], v = [1, 2..., h]. \end{aligned} \tag{1}$$

where w and h is the width and height of a video. Apart from the different input, temporal stream holds the same structure and layer parameters with spatial stream as we can see in Table 1.

Fusion For a given video frame I and its optical flow feilds I_{fl}, each stream performs feed forward and computes softmax scores respectively. The final score is then fusion by a linear SVM.

Pre-training Before training, we first pre-compute the optical flow displacement fields using the method of [1] whose GPU implementation can be obtained from OPENCV toolkit. Then, we fine-tune the spatial stream on self-created smoke dataset after pre-training

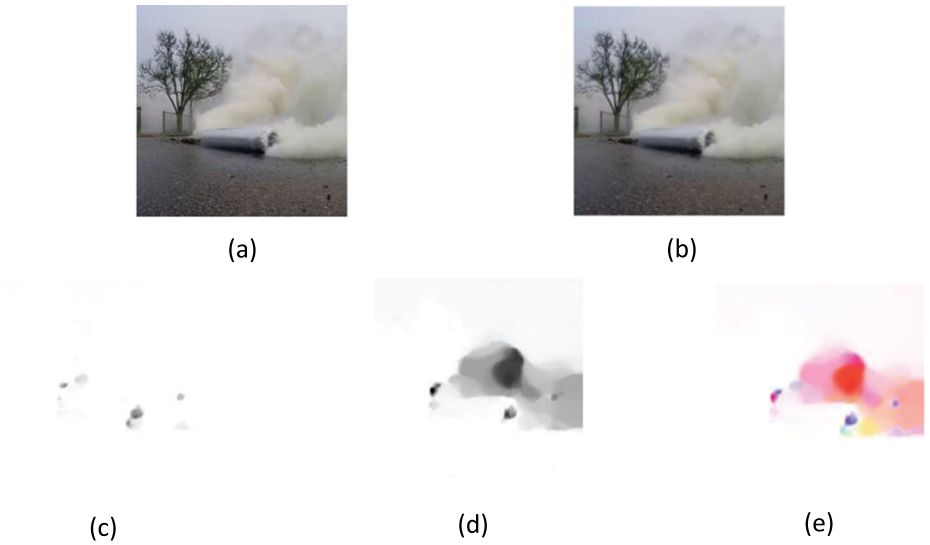

Fig. 4 Example of optical flow displacement field

on ImageNet [21]. Similar to spatial stream, temporal stream is pre-trained on UCF101 dataset [15](we select the neighbor key frames in video clips and compute the optical flow as pre-training samples.).

3.2 Proposal 2: enhanced spatial-temporal ConvNet for video smoke detection

The architecture of proposal 1 can effectively capture appearance and motion information but remains to be improved. In general, there are two shortcomings: (1) optical flow pre-computation is time-consuming and even much slower than ConvNet feed forward; (2) spatial stream and temporal stream is just combined in final classification but separately learns features in training process(can not regularize each other).

To address the above shortcomings, we propose an enhanced spatial-temporal architecture which can avoid complex optical flow pre-computation. Additionally, the enhanced deep architecture takes two neighbor frames as input and can jointly extract intra-frame appearance feature and inter-frame motion feature by multi-task learning strategy. An illustration of the enhanced deep architecture can be seen in Fig. 5 and the parameters of each layer are listed in Table 2.

Enhanced deep architecture Given neighbor two frames of a video sequence, we concatenate them as the architecture's input with six channels of size 227×227. The enhanced architecture contains five convolutional layers, three fully connected layers and four deconvolutional layers. Five convolutional layers have the same parameters as the architecture in proposal 1 which can reduce the input to 256 feature maps with size of 6×6. For classification, fully connected layers follow after the convolutions and a softmax classifier can recognize smoke or non-smoke. Deconvolutional networks are also connected after the convolutional layers which can upsample the last feature maps and output the optical flow estimation with the size of $96 \times 96 \times 2$. From another viewpoint, the enhanced architecture is learned by two supervisions. One is softmax loss for smoke recognition, another is euclidean

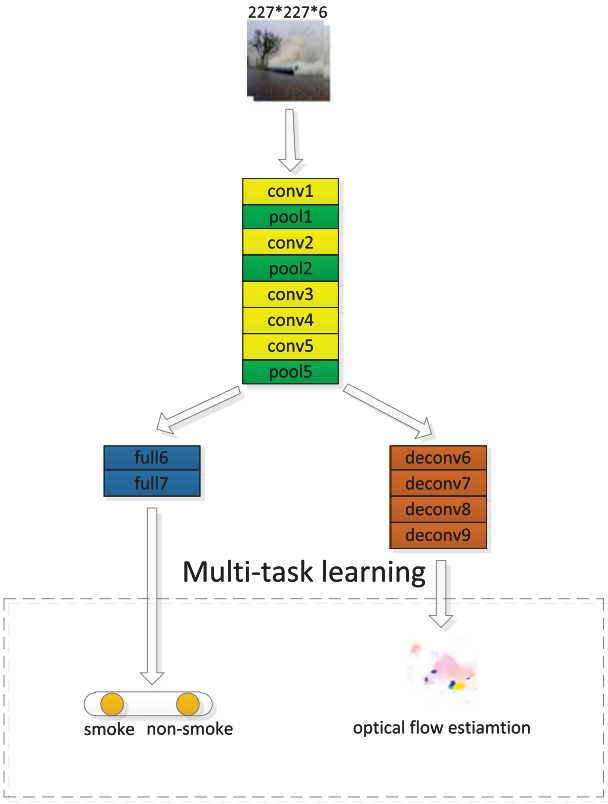

Fig. 5 Architecture of proposal 2: enhanced spatial-temporal stream ConvNet

Table 2 The relevant layers and their parameters of proposals 2

Layer	Layer Type	Parameter
input	input	Image size: 227×227×6
conv1	convolution	Filter size: 11×11, Filter number: 96, Stride:4
pool1	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv2	convolution	Filter size: 5×5, Filter number: 256, Stride:1
pool2	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
conv3	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv4	convolution	Filter size: 3×3, Filter number: 384, Stride:1
conv5	convolution	Filter size: 3×3, Filter number: 256, Stride:1
pool5	pooling	Pooling Method:Max, Kernel size: 3×3, Stride:2
full6	fully-connected	Neurons output: 4096
full7	fully-connected	Neurons output: 2048
deconv6	deconv	Filter size: 3×3, Filter number:256, Stride: 2
deconv7	deconv	Filter size: 3×3, Filter number:128, Stride: 2
deconv8	deconv	Filter size: 3×3, Filter number:64, Stride: 2
deconv9	deconv	Filter size: 3×3, Filter number:2, Stride: 2

loss to measure the distance between the optical flow estimations and the groundtruth. The combination of two supervisions can regularize the ConvNet to joint learn appearance and motion. Moreover, for testing, the enhanced architecture is much faster than the architecture of proposal 1 which accelerates the smoke detection to real time.

Deconvolution Deconvolutional layers have been widely used in semantic segmentation [18, 20], image restoration [19] and motion estimation [3, 12]. Derived from convolution operation, deconvolution can refine the coarse feature maps through convolution-like operations with multiple learned filters. However, deconvolution is a one-to-multi mapping operation which makes the feature maps larger than before. To estimate the optical flow, in our application, we utilize the deconvolution operation to the feature maps and concatenate them with corresponding feature maps of the convolutional layers whose implementaion is similar to the reference of [3]. After four times deconvoltion, the architecture outputs the estimated optical flow with size of 96×96×2 which represents the horizontal component and vertical component of the motion, respectively.

Multi-task learning Unlike the architecture of proposal 1, which trains spatial and temporal stream separately, the enhanced architecture trains for recognizing smoke and estimating optical flow simultaneously through utilizing multi-task learning strategy. We can define training set $\mathbf{X} = \{\chi^1, \chi^2, \chi^3, ..., \chi^K\}$, χ^k=(I^k, O^k_{fl}, l^k),where χ^k is the k-th training sample, O^k_{fl} represents its groundtruth optical flow, l^k is its class label(smoke or non-smoke), I^k is the six channels input concatenated by two consecutive neighbor frames. Convolutional layers can hierarchically learn feature representations and can be denoted as:

$$\mathbf{F}^k = Conv(I^k|\theta_{conv}), \tag{2}$$

where $\mathbf{F}^k$ is the learned feature maps of the k-th samples, $Conv(\cdot|\theta_{conv})$ is the convolution operation, and θ_{conv} denotes the parameters of convolutional layers.

Then, we learn the architecture with two supervisions. The former is the softmax loss for smoke recognition.

$$\mathbf{f}^k = FC(\mathbf{F}^k|\theta_{fc}) = \theta^T_{fc}\mathbf{F}^k, \tag{3}$$

$$\mathcal{L}_{cls}(\mathbf{f}^k, l^k, \theta_{cls}) = -\left[\sum_{i=1}^{n} 1\{i = l^k\} \log P(i = l^k|\mathbf{f}^k, \theta_{cls})\right], \tag{4}$$

where $FC(\cdot|\theta_{fc})$ represents the operation of fully connected layers which is essentially the inner product of the parameters θ_{fc} and the feature maps $\mathbf{F}^k$. $\mathbf{f}^k$ is both the output of the fully connected layers and the input of the softmax classifier. l^k is the class label and θ_{cls} represents the parameters of softmax classifier. 1{·} is the indicator function. 1{a true statement}=1 , and 1{a false statement}=0. P(·|$\mathbf{f}^k$, θ_{cls}) represents the predicted probability distribution.

The later is the euclidean loss which measures the distance between optical flow estimations and the groundtruth.

$$\hat{O}^k_{fl} = Deconv(\mathbf{F}^k|\theta_{dec}), \tag{5}$$

$$\mathcal{D}\left(O^k_{fl}, \hat{O}^k_{fl}\right) = \frac{1}{2}||O^k_{fl} - \hat{O}^k_{fl}||^2_2, \tag{6}$$

where $Deconv(\cdot|\theta_{dec})$ is the deconvolution operation, θ_{dec} denotes the parameter of deconvolutional layers, $\hat{O}^k_{fl}$ represents the estimated optical flow, and we use the L2-norm to measure the distance between estimation and groundtruth.

Therefore, the smoke detection task can be simplified to compute the optima θ_{conv}, θ_{fc}, θ_{cls} and θ_{dec} under the combined loss function as followed:

$$\begin{aligned}\theta^*_{conv}, \theta^*_{fc}, \theta^*_{cls}, \theta^*_{dec} = \operatorname{argmin}_{\theta_{conv},\theta_{fc},\theta_{cls},\theta_{dec}} \left(\sum_{k=1}^{N} \left(\lambda_c \mathcal{L}_{cls} + \lambda_o \mathcal{D}\left(O^k_{fl}, \hat{O}^k_{fl}\right)\right)\right. \\ \left. + \|\theta_{conv}\|_2^2 + \|\theta_{fc}\|_2^2 + \|\theta_{cls}\|_2^2 + \|\theta_{dec}\|_2^2\right),\end{aligned} \tag{7}$$

where θ_{conv} denotes the convolutional layers parameters, θ_{fc} denotes the fully connected layers parameters, θ_{cls} is the parameters of softmax classifier, θ_{dec} represents the deconvolutional parameters. In right equation, $\mathcal{L}_{cls}$ is the softmax loss, $\mathcal{D}$ denotes the euclidean loss, while λ_c and λ_o are hyperparameters which can balance the smoke recognition and optical flow estimation.

The parameters of each layer are optimized by stochastic gradient decent, and the multi-task learning algorithm is summarized in Algorithm 1. The architecture jointly learns class and optical flow during training, but it should be noted that, in testing stage, the deconvolution operation can be ignored so as to speed up the ConvNet feed forward.

Algorithm 1 The Smoke Detection Learning Algorithm

Input: training set $\chi = \{I^i, O^i_{fl}, l^i\}$, initialized parameters θ_{conv}, θ_{fc}, θ_{cls} and θ_{dec}, hyperparameter λ_c, λ_o, learning rate $\eta(t)$, $t \leftarrow 0$, times of iteration N, batch size M.

While $t \neq N$ **do**

t←t+1 sample M training samples $\{I^i, O^i_{fl}, l^i\}$ from χ

$\mathbf{F}^i = Conv(I_i, \theta_{conv})$, $\mathbf{f}^i = FC(\mathbf{F}^i, \theta_{fc})$, $\hat{O}^i_{fl} = Deconv(\mathbf{F}^i, \theta_{dec})$

$\nabla\theta_{cls} = \sum_{i=1}^{M} \frac{\partial \mathcal{L}_{cls}(\mathbf{f}^i, l_i, \theta_{cls})}{\partial \theta_{cls}}$

$\nabla\mathbf{f}^i = \lambda_c \cdot \frac{\partial \mathcal{L}_{cls}(\mathbf{f}^i, l_i, \theta_{cls})}{\partial \mathbf{f}^i}$

$\nabla\theta_{fc} = \sum_{i=1}^{M} \nabla\mathbf{f}^i \cdot \frac{\partial FC(\mathbf{F}^i, \theta_{fc})}{\partial \theta_{fc}}$

$\nabla\hat{O}^i_{fl} = \lambda_o \cdot \frac{\partial \mathcal{D}(O^i_{fl}, \hat{O}^i_{fl})}{\partial \hat{O}^i_{fl}}$

$\nabla\theta_{dec} = \sum_{i=1}^{M} \nabla\hat{O}^i_{fl} \cdot \frac{\partial Deconv(\mathbf{F}^i, \theta_{dec})}{\partial \theta_{dec}}$

$\nabla\mathbf{F}^i = \frac{\partial FC(\mathbf{F}^i, \theta_{fc})}{\partial \mathbf{F}^i} + \frac{\partial Deconv(\mathbf{F}^i, \theta_{dec})}{\partial \mathbf{F}^i}$

$\nabla\theta_{conv} = \sum_{i=1}^{M} \nabla\mathbf{F}^i \cdot \frac{\partial Conv(I_i, \theta_{conv})}{\partial \theta_{conv}}$

update $\theta_{cls} = \theta_{cls} - \eta(t) \cdot \nabla\theta_{cls}$, $\theta_{fc} = \theta_{fc} - \eta(t) \cdot \nabla\theta_{fc}$, and $\theta_{dec} = \theta_{dec} - \eta(t) \cdot \nabla\theta_{dec}$, $\theta_{conv} = \theta_{conv} - \eta(t) \cdot \nabla\theta_{conv}$

End while

Output $\theta_{conv}, \theta_{fc}, \theta_{cls}$

Pre-training It is impossible to manually annotate the groundtruth optical flow O^k_{fl} from the smoke video, so we use the method of [1] to generate the pseudo groundtruth label. We first pre-train the convolutional layers and fully connected layers on ImageNet [21] (Two identical images are concatenated as one pre-training sample with size of 227×227×6) by using softmax loss as supervision, then fine-tune the whole architecture with the joint supervision of softmax loss and euclidean loss.

Compared with the proposal 1, the advantages of the enhanced architecture can be summarized as follow:

- The enhanced architecture jointly learns class and optiacl flow with two supervisions which can regularize each other and reduce over-fitting in training process.
- In testing stage, we just concatenate two neighbor frames as input without optical flow pre-computation, which can accelerate smoke detection to real time.

4 Experiment

We use the open source toolbox Caffe [13] to implement the proposed spatial-temporal architecture and some modifications are applied. For experiment, we perform on a workstation with Intel Core I7 , NVIDIA GTX TITAN X GPU, and the operating system of Ubuntu 16.04. We utilize the Stochastic Gradient Decent(SGD) to update the parameters of each layer with the mini-batch size of 80, momentum of 0.9 and initial learning rate of 0.01. The training procedure maintains until the validation accuracy remains unchanged for ten consecutive epoches.

4.1 Experiments setup

Self-created dataset Existing smoke detection dataset lacks challenging negative non-smoke videos, such as cloud, fog or haze. For this reason, we create our own dataset containing 157 videos(71 smoke videos & 86 non-smoke videos) which are all searched from Internet engines, totally 2.5 hours of recording. Then, we truncate the source videos to 3738 clips(1786 positive & 1304 simple negative & 648 challenge negative), with average 60 frames per video clips. Table 3 shows the total number of frames in our dataset, where some non-smoke videos present subtle intra-class variation as we can see in Fig. 1.

Data augmentation It has been proven that small training samples may lead to serious over-fitting [9]. Thus, in this paper, we adopt two data augmentation strategies. One is horizontal reflection and rotation, Another is random crop (we first rescale frames to the size of 256$\times$ 256 and then randomly select crops with the size of 227$\times$227). Adding augmented data makes our method invariant to transformation and effectively increases the robustness of the architecture.

Evaluation criteria In order to quantify the comparative experiment results, we use three evaluation criteria: Detection Rate(DR), False Alarm Rate(FAR) and Accuracy Rate(AR), which can be denoted as:

$$DR = \frac{P_p}{P_p + P_n} \times 100\%, \tag{8}$$

Table 3 Total smoke and non-smoke samples in our dataset

Video type	Number of videos	Number of clips	Number of frames
Smoke video	71	1786	107584
Simple non-smoke video	55	1304	77973
Challenge non-smoke video	31	648	38965

$$FAR = \frac{N_p}{N_p + N_n} \times 100\%, \tag{9}$$

$$AR = \frac{P_p + N_n}{P_p + P_n + N_p + N_n} \times 100\%, \tag{10}$$

where P_p is the true positive, P_n is the false negative, N_p is the false positive, and N_n is the true negative. Our sole aim is to achieve high detection rate(DR), high accuracy rate(AR) and low false alarm rate(FAR) at the same time.

Cross validation For experiments, we randomly divide our video sequences into 5 sets with the same proportion of positive and negative, and then perform 5 fold cross validation on our own dataset. The comparison with related methods can be seen in Table 4.

4.2 Comparison with handcrafted feature-based methods

We investigate the performance of some handcrafted feature-based methods. The quantitative results are listed in Table 4.

In [24], the authors combined static Gray level Co-occurrence Matrix(GLCM) features and global motion features, and then utilized BP neural network for classification. We test their methods on our datasets and the accuracy rate is 90.4%. In [27], Tian et al. utilized the texture descriptors NRLBP as features and then trained a SVM classifier to determine smoke or non-smoke. We repeated their method which achieved the accuracy rate of 91.3%. Ko et al. proposed a solution based on bag of words(BOW) features and their implementation performs better than the above two methods, achieving the accuracy rate of 92.5%.

Although handcrafted feature-based methods have ever been widely used in the task of video smoke detection, they apparently fall behind the deep learning based methods of all criteria as we can see in Table 4.

4.3 Comparison with deep learning based method

Here, we report the experiment results of our proposed methods and comparisons with other deep learning methods.

Tao et al. [26] applied AlexNet to extract appearance features and recognize smoke of each frame, since their implementations can be regarded as a single spatial stream ConvNet, which achieves the DR of 95.2%, FAR of 6.2%, and AR of 94.4% in our dataset. Yin et al. [31] proposed a batch normalization strategy to improve the generalization ability of

Table 4 Detection Rate(DR), False Alarm Rate(FAR) and Accuracy Rate(AR) of our proposed method, and comparisons with handcrafted feature-based methods and other deep learning based methods

Algorithms	DR	FAR	AR
GLCM+NN (Srisuwan et al. [24])	91.2%	10.3%	90.4%
NRLBP+SVM (Tian et al. [27])	92.1%	9.3%	91.3%
BoW+RF (Ko et al. [16])	93.9%	8.7%	92.5%
Spatial stream ConvNet (AlexNet, Tao et al. [26])	95.2%	6.2%	94.4%
Deep normalization ConvNet (Yin et al. [31])	95.9%	5.9%	95.0%
Temporal stream ConvNet	92.7%	7.9%	92.4%
Spatial-temporal stream ConvNet(Proposal 1)	95.4%	5.8%	94.8%
Enhanced spatial-temporal ConvNet(Proposal 2)	97.0%	3.5%	96.7%

the ConvNet. We repeated their implementations in our dataset, and boosted the accuracy rate(AR) to 95.0% .

The architecture of proposal 1 combines spatial stream and temporal stream by late fusion, obtaining the DR of 95.4%, FAR of 5.8%, AR of 94.8%, which is worse than deep normalization ConvNet and just slightly better than AlexNet. The enhanced architecture of proposal 2 utilizes a multi-task learning strategy to jointly recognize smoke and estimate optical flow by computing frame difference similar to the reference of [39], achieving the state of the art performance with the DR of 97.0%, FAR of 3.5%, AR of 96.7%. Contrary to spatial stream, temporal stream ConvNet takes the optical flow as input and extracts inter-frame motion features for classification, achieving the DR of 92.7%, FAR of 7.9%, and AR of 92.4%.

In summary, our proposed methods have better performance than AlexNet and deep normalization ConvNet(single spatial stream ConvNet), due to the combination of appearance and motion information. Further more, we can observe that multi-task learning strategy utilized in proposal 2 compares favourably to that combined with late SVM fusion in proposal 1.

4.4 Ablation studies for multi-task learning architecture

The architecture of proposal 2 adopts multi-task learning strategy to learn class and estimate optical flow. λ_c and λ_o are two key hyperparameters in the multi-task architecture which weight the softmax loss for smoke recognition and the euclidean loss for optical flow estimation. To further illustrate the superiority of multi-task learning architecture, we run an ablation analysis to evaluate the performance of softmax supervision, euclidean supervision, and the joint multi-task supervision.

Here, we utilizes Accuracy Rate(AR) and End-Point-Error(EPE) to evaluate the standard of smoke recognition and optical flow estimation respectively. End-Point-Error(EPE) is the sum of L2-norm distance between the groundtruth of optical flow and the estimated flow over all pixels which can be denoted as:

$$EPE = ||O_{fl} - \hat{O}_{fl}||_2, \tag{11}$$

where O_{fl} is the optical flow groundtruth and $\hat{O}_{fl}$ represents the estimation.

We report the experiment results with three different groups of parameters: removing the euclidean supervision($\lambda_c = 1, \lambda_o = 0$), removing the softmax supervision($\lambda_c = 0, \lambda_o = 1$) and the joint multi-task supervision($\lambda_c = 1, \lambda_o = 0.1$). The quantitative performance is listed in Table 5.

We can observe that the multi-task architecture achieves higher AR and lower EPE, compared with the single-task architecture. Then our goal is converted to select an appropriate group of parameters for smoke recognition.

Table 5 Accuracy Rate(AR) and End-Point-Error(EPE) in the ablation analysis: softmax supervision, euclidean supervision and the multi-task supervision

Algorithms	AR	EPE
Softmax supervision($\lambda_c = 1, \lambda_o = 0$)	94.4%	–
Euclidean supervision($\lambda_c = 0, \lambda_o = 1$)	–	12.3
Multi-task supervison($\lambda_c = 1, \lambda_o = 0.1$)	95.5%	11.7

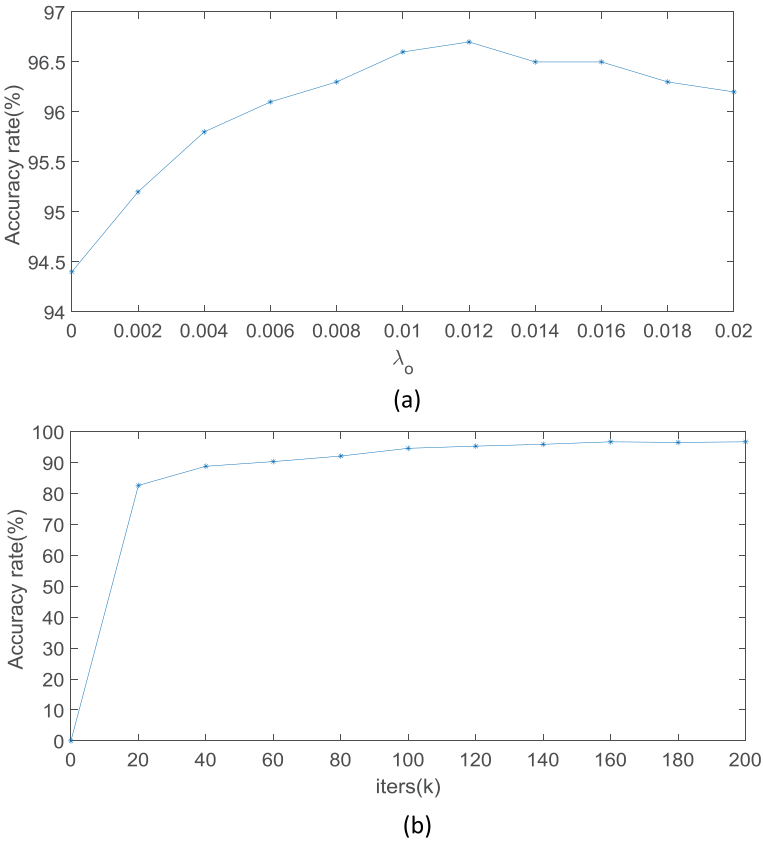

Fig. 6 Line charts of accuracy rate

Here, λ_c is set to be 1 and we conduct an experiment to find the optima λ_o. As we can see in Fig. 6a, λ_o ranges from 0 to 0.02 by the rate of 0.002. While λ_o is set to be 0.012, it achieves the highest accuracy rate; that is to say, 0.012 is the best reference in our application. In addition, we illustrate the convergence process as we can see in Fig. 6b.

4.5 Comparison of the processing speed

In Table 6, we evaluate the processing speed of our proposed methods using the criterion of frame per second(fps). The architecture of proposal 1 performs nearly 25 times slower

Table 6 The processing speed of our proposed method, and comparison with other deep learning based methods

Algorithms	fps
AlexNet (Tao et al. [26])	264
Spatial-temporal stream ConvNet(Proposal 1)	11
Enhanced spatial-temporal ConvNet(Proposal 2)	196

than AlexNet, since it consumes so much time in optical flow pre-computation; while the enhanced architecture of proposal 2 skips the pre-computation and achieves the frame rate of 196 fps, which realizes the real-time smoke detection.

5 Conclusion

In this paper, we present an enhanced spatial-temporal convolutional neural network for video smoke detection, which utilizes a multitask learning method to jointly capture intra-frame appearance information and inter-frame motion information. Experiments our self-created dataset show that our proposed method achieves highest detection rate and lowest false alarm rate(state-of-the-art) in a speed of very faster than real time. For future researches, how to extract motion features more accurately and completely may be a potential research interest.

Acknowledgements The authors would like to thank the editor and the anonymous reviewers for their valuable comments and constructive suggestions. This work was supported by the National Key Science & Technology Pillar Program of China (No. 2014BAG01B03), the National Natural Science Foundation of China (No. 61374194), Key Research and Development Program of Jiangsu Province (No. BE2016739), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. Computer vision - ECCV 2004: 8th European conference on computer vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV, pp 25–36
2. da Penha OS, Nakamura EF (2010) Fusing light and temperature data for fire detection. In: The IEEE Symposium on computers and communications, pp 107–112. https://doi.org/10.1109/ISCC.2010.5546519
3. Dosovitskiy A, Fischery P, Ilg E, Hausser P, Hazirbas C, Golkov V, Smagt VD, Cremers P, Brox D, Flownet T (2015) Learning optical flow with convolutional networks. In: 2015 IEEE International conference on computer vision (ICCV), pp 2758–2766. https://doi.org/10.1109/ICCV.2015.316
4. Frizzi S, Kaabi R, Bouchouicha M, Ginoux JM, Moreau E, Fnaiech F (2016) Convolutional neural network for video fire and smoke detection. In: IECON 2016 - 42nd Annual conference of the IEEE industrial electronics society, pp 877–882. https://doi.org/10.1109/IECON.2016.7793196
5. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on computer vision and pattern recognition, pp 580–587
6. Gubbi J, Marusic S, Palaniswami M (2009) Smoke detection in video using wavelets and support vector machines. Fire Safe J 44(8):1110–1115. https://doi.org/10.1016/j.firesaf.2009.08.003. http://www.sciencedirect.com/science/article/pii/S0379711209001155
7. Han Y, Yang Y, Wu F, Hong R (2015) Compact and discriminative descriptor inference using multi-cues. IEEE Trans Image Process 24(12):5114–5126. https://doi.org/10.1109/TIP.2015.2479917
8. Han J, Zhang D, Cheng G, Liu N, Xu D (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100. https://doi.org/10.1109/MSP.2017.2749125
9. Howard AG (2013) Some improvements on deep convolutional neural network based image classification. CoRR abs/1312.5402
10. Hu Y, Chang H, Nian F, Wang Y, Li T (2016) Dense crowd counting from still images with convolutional neural networks. J Vis Commun Image Represent 38:530–539. https://doi.org/10.1016/j.jvcir.2016.03.021. http://www.sciencedirect.com/science/article/pii/S1047320316300256
11. Huang X (2018) Automatic video superimposed text detection based on nonsubsampled contourlet transform. Multimed Tools Appl 77(6):7033–7049. https://doi.org/10.1007/s11042-017-4619-8

12. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2016) Flownet 2.0: evolution of optical flow estimation with deep networks. CoRR abs/1612.01925
13. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: MM 2014 - Proceedings of the 2014 ACM conference on multimedia
14. Kaiser T (2000) Fire detection with temperature sensor arrays. In: Proceedings IEEE 34th annual 2000 international carnahan conference on security technology (Cat. No.00CH37083), pp 262–268. https://doi.org/10.1109/CCST.2000.891198
15. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on computer vision and pattern recognition, pp 1725–1732. https://doi.org/10.1109/CVPR.2014.223
16. Ko B, Park J, Nam JY (2013) Spatiotemporal bag-of-features for early wildfire smoke detection. Image Vis Comput 31(10):786–795. https://doi.org/10.1016/j.imavis.2013.08.001
17. Krizhevsky A, Ilya S, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
18. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
19. Mao X, Shen C, Yang Y (2016) Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections. CoRR abs/1603.09056
20. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: 2015 IEEE International conference on computer vision (ICCV), pp 1520–1528. https://doi.org/10.1109/ICCV.2015.178
21. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
22. Sainath T, Kingsbury B, Mohamed A, Dahl GE, Saon G, Soltau H, Beran T, Aravkin AY, Ramabhadran B (2013) Improvements to deep convolutional neural networks for lvcsr. In: IEEE Workshop on automatic speech recognition and understanding, pp 315–320
23. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Adv Neural Inf Process Syst, 1
24. Srisuwan T, Ruchanurucks M (2013) Smoke detection using glcm, wavelet, and motion. In: Proceedings of SPIE - the international society for optical engineering, p 9069
25. Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: IEEE Conference on computer vision and pattern recognition, pp 1891–1898
26. Tao C, Zhang J, Wang P (2016) Smoke detection based on deep convolutional neural networks. In: 2016 International conference on industrial informatics - computing technology, intelligent technology, industrial information integration (ICIICII), pp 150–153. https://doi.org/10.1109/ICIICII.2016.0045
27. Tian H, Li W, Ogunbona P, Nguyen DT, Zhan C (2011) Smoke detection in videos using non-redundant local binary pattern-based features. In: 2011 IEEE 13th International workshop on multimedia signal processing, pp 1–4. https://doi.org/10.1109/MMSP.2011.6093844
28. Toreyin B, Dedeolu Y, Enis A, Etin C (2005) Wavelet based real-time smoke detection in video. In: Proceedings of 13th European signal processing conference
29. Xu G, Zhang Y, Zhang Q, Lin G, Wang J (2017) Domain adaptation from synthesis to reality in single-model detector for video smoke detection. arXiv:1709.08142
30. Yao X, Han J, Zhang D, Nie F (2017) Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. IEEE Trans Image Process 26(7):3196–3209. https://doi.org/10.1109/TIP.2017.2694222
31. Yin Z, Wan B, Yuan F, Xia X, Shi J (2017) A deep normalization and convolutional neural network for image smoke detection. IEEE Access 5:18,429–18,438. https://doi.org/10.1109/ACCESS.2017.2747399
32. Yuan F (2008) A fast accumulative motion orientation model based on integral image for video smoke detection. Pattern Recogn Lett 29(7):925–932. https://doi.org/10.1016/j.patrec.2008.01.013. http://www.sciencedirect.com/science/article/pii/S0167865508000263
33. Yuan F (2011) Video-based smoke detection with histogram sequence of lbp and lbpv pyramids. Fire Safety J 46(3):132–139. https://doi.org/10.1016/j.firesaf.2011.01.001. http://www.sciencedirect.com/science/article/pii/S0379711211000026
34. Yuan F (2012) A double mapping framework for extraction of shape-invariant features based on multi-scale partitions with adaboost for video smoke detection. Pattern Recogn 45(12):4326–4336. https://doi.org/10.1016/j.patcog.2012.06.008. http://www.sciencedirect.com/science/article/pii/S0031320312002786

35. Yuan F, Shi J, Xia X, Fang Y, Fang Z, Mei T (2016) High-order local ternary patterns with locality preserving projection for smoke detection and image classification. Inf Sci 372:225–240. https://doi.org/10.1016/j.ins.2016.08.040. http://www.sciencedirect.com/science/article/pii/S0020025516306168
36. Zeiler M, Fergus R (2014) Visualizing and understanding convolutional networks. In: Europe Conference on computer vision, pp 818–833
37. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 833–841. https://doi.org/10.1109/CVPR.2015.7298684
38. Zhang Q, Xu J, Xu L, Guo H (2016) Deep convolutional neural networks for forest fire detection. In: International forum on management, education & information technology application
39. Zhao S, Liu Y, Han Y, Hong R, Hu Q, Tian Q (2017) Pooling the convolutional layers in deep convnets for video action recognition. IEEE Trans Circ Syst Vid Technol PP(99):1–1. https://doi.org/10.1109/TCSVT.2017.2682196

Yaocong Hu received the B.S. degree in automation from Anhui Polytechnic University, Wuhu, China, in 2014 and received the M.S. degree in pattern recognition and intelligent system from Anhui University, Hefei, China, in 2017. Now, he is currently working toward the Ph.D. degree with the School of Automation, Southeast University. His current research interests include image processing and deep learning.

Xiaobo Lu received the B. S. degree in Department of Precision Instruments from Shanghai Jiao Tong University, Shanghai, China, the M.S. degree in School of Automation from Southeast University, Nanjing, China, the Ph. D. degree in Department of Testing Engineering from Nanjing University of Aeronautics and Astronautics and he did his postdoctoral research at Chien-Shiung Wu Laboratory at Southeast University from 1998 to 2000.

Now, he is a professor at the School of Automation and the deputy director of the Detection Technology and Automation Research Institute in Southeast University. He is a coauthor of the book An Introduction to the Intelligent Transportation Systems (Beijing: China Communications Press, 2008). He has earned many research awards, such as the first prize in Natural Science Award of the Ministry of Education of China and the prize in Science and Technology Award of Jiangsu province. His research interests include image processing, signal processing, pattern recognition, and computer vision.