Multimedia Tools and Applications (2021) 80:27407–27421
https://doi.org/10.1007/s11042-021-11037-1

Fire smoke detection based on target-awareness and depthwise convolutions

Yunji Zhao[1] · Haibo Zhang[2] · Xinliang Zhang[1] · Xiangjun Chen[3]

Received: 29 June 2020 / Revised: 24 February 2021 / Accepted: 5 May 2021 /
Published online: 18 May 2021

Abstract
Because smoke usually appears before a flame arises, fire smoke detection is significant for early warning systems. This paper proposes a TADS (Target-awareness and Depthwise Separability) algorithm based on target-awareness and depthwise separability. Current deep learning methods with pre-trained convolutional neural networks by abundant and vast datasets are used to realize generic object recognition tasks. As for smoke detection, collecting large quantities of smoke data is challenging for small sample smoke objects. The basis is that the objects of interest can be arbitrary object classes with arbitrary forms. Thus, deep feature maps acquired by target-awareness pre-trained networks are used for modeling these objects of arbitrary forms to distinguish them from unpredictable and complex environments. The authors introduced this scheme to deal with smoke detection. The depthwise separable method with a fixed convolution kernel replacing the training iterations can improve the algorithm's speed to meet the enhanced requirements of real-time fire spreading for detecting speed. The experimental results demonstrate that the proposed algorithm can detect early smoke in real-time, and it is superior to the state-of-the-art methods in terms of accuracy and speed.

Keywords Fire smoke detection · Target-awareness · Depthwise separable · Fixed convolution kernel · TADS

✉ Yunji Zhao
auyjz@hpu.edu.cn

[1] School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000 Henan, China

[2] School of Big Data and Artificial Intelligence, Xinyang University, Xinyang 464000 Henan, China

[3] College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000 Henan, China

1 Introduction

Image pressing techniques are widely used in pattern recognition and artificial intelligence, such as object tracking [4], image classification [7], object detection, and so on. Intelligent video surveillance has become a research focus on computer vision with the development of computer science. Smoke detection is a promising method for fire alarm systems, especially in wide-open forest environments. Automatic fire detection systems play an essential role in the early detection and response of unpredictable scenes. Ideal performance concerning smoke detection and analysis tasks are challenging because of the multiformity of form, swing, changing smoke color tones, environmental illumination, and low-resolution images of forest scenes. Traditional video smoke detection methods are mostly based on pattern recognition, and digital image processing techniques depend on conventional image processing techniques [5], such as texture [36], wavelets [31], spatial features [1, 10]. Park et al. [22] utilized motion detection and support vector machines to detect the smoke in the engine room of the ship. According to Morerio et al. [19], the CIELab color space was used to perform a smoke chromatic feature clustering method to analyze smoke color features. Tian et al. used a histogram of oriented gradient (HOG) descriptors to extract smoke spatial features [30]. Xiong et al. used the adaptive Gaussian mixture model (GMM) to construct the background model [32]. The values that do not match background Gaussian pixels were grouped as moving blobs using the connected component analysis to detect smoke. Ye et al. detected smoke and flame based on the color features and wavelet analysis [36]. Luo et al. used background dynamic update and dark channel priori to detect the smoke regions [16]. Subsequently, CNN was applied to extract the semantic features and determine whether there is fire in the suspected region. Park et al. selected frames to construct a temporal fire-tube and extracted a histogram of the fire tube to describe fire's night-time characteristics [23]. YOLOv3 was used to detect the fire region. The traditional features of smoke are easily affected by the illumination changes, complex background, scale changes, and so on. Compared with conventional smoke features, the semantic features have good robustness and stability.

In recent years, many machine visions tasks had made significant progress concerning the applications of realistic scenarios and public benchmark datasets by deep learning approaches [3, 15, 16, 21, 23, 25, 27]. Video smoke detection using a relatively deep network has attracted many researchers. Smoke detection methods based on deep learning adopt the mainstream deep learning framework. Saeed et al. used an Adaboost-MLP model to predict the fire in the videos and used the convolutional neural networks to predict the fire in the images [27]. The videos and images were captured from the cameras installed for the surveillance. Pundir et al. utilized one deep learning framework to extract the image-based features of the smoke patches, to describe the smoke-color, smoke-texture; and the other deep learning framework to extract motion-based features [25]. Combined features were used to train the deep CNN to realize the smoke classification. Lin et al. investigated the combined framework of faster RCNN and 3D CNN [15]. The RCNN with non-maximum annexation was used to locate the smoke, and 3D CNN combined with dynamic spatial–temporal information was used to recognize smoke. According to Yin et al. [37], the normalization and convolutional neural network (NCNN) was applied to detect smoke in the smoke video. According to Mao et al., a multichannel convolutional neural network was proposed to extract deep features for fire detection [18]. Sharma et al. used two pretrained convolutional neural networks (CNNs), VGG16 and Resnet50, to detect early fires [29]. Muhammad et al. proposed a cost-effective CNN to balance complex computations and accuracy [20]. Xu et al. applied synthetic smoke images to deal with a lack of training data for

CNN [33]. According to Pundir et al. [24], a background subtraction algorithm was proposed to preprocess smoke video to significantly display smoke areas, and a deep belief network is used to classify smoke. Based on the proposal of interest, the deep learning framework is a class of CNN architectures combined with a region proposal method. The region-based CNN (RCNN) [6] is a CNN extension combined with selective search to detect objects. A region proposal network (RPN) is added to a typical CNN to anchor the object regions of interest. Faster R-CNN [26] was proposed for pretrained VGG16 combined with RPN to classify objects and regress bounding boxes. According to Lin et al. [14], a Faster R-CNN was adopted to extract smoke areas crudely, and a 3D-CNN was used to classify smoke videos. The smoke detection algorithms use deep features as the description of the template. Deep features extraction and classification scheme affect the speed of detection. The pretrained deep model is always used to extract the smoke's semantic features and classify the smoke from the image. The deep features are not all useful for describing the smoke. The redundant deep features may affect not only the accuracy but also the efficiency of the smoke detection.

The deep saliency network for smoke detection is a novel method that aims to emphasize the most important object regions in video frames. According to Xu et al. [35], salient convolutional neural networks based on pixel-level and object-level were used to extract smoke saliency map. Jia et al. applied a saliency detection model to segment the smoke regions based on the pixels and motion features [11]. In this paper, an end-to-end framework for video smoke detection is proposed. In the framework of the correlation filter, deep features extracted by CNN are processed by target-awareness to realize dimension reduction. To meet smoke detection's real-time requirements, a depthwise scheme with a fixed convolution kernel is applied to replace the traditional convolution. In the response image, the maximum value is used to predict the position of the detection area. A multiscale scheme can be used to determine the rectangle of the smoke area. This paper is organized as follows: Section 1 reviews the related works. In Section 2, TADS is introduced. The experimental results are presented in Section 3, and the conclusion is presented in Section 4.

2 Related work

Smoke detection based on deep learning methods is different from traditional image processing methods. The deep learning algorithms can extract multiclass features which are not limited to one or two typical image processing features. Yuan et al. [38] used fully convolutional networks (FCNs) to realize semantic segmentation. A deep smoke segmentation network was also proposed to segment blurry smoke images via training high-quality segmentation masks. Traditional vision-based smoke detection methods [5] always divide each video frame into blocks and extract stable features in each block to classify smoke or nonsmoke. These methods' highlighted performance relies on robust visual object forms to distinguish smoke from video scenes with clear background differentiation. However, fires are always accompanied by complex background effects and fuzzy real-time video data, which can hardly supply high-quality and high-contrast videos. Existing technical conditions cannot strictly meet video detection requirements for large quantities of data for small sample objects. Xu et al. proposed synthetic smoke images to meet dataset requirements [34]. However, in visual detection, the objects of interest can be arbitrary object classes with arbitrary forms. Therefore, it is impossible to complete all realistic scenarios. As a result, deep feature maps for pretraining are weak in modeling these objects of arbitrary forms for distinguishing them from unpredictable and complex environments.

In this paper, according to the target-awareness deep tracking (TADT) algorithm [13], TADS is proposed with a target-awareness strategy to select useful deep features for object representation. Target-awareness is realized according to regression loss. According to Maaten et al. [17], the T-SNE model shows the difference between target-awareness features and original features. Pretrained deep features are less effective than target-awareness deep features for discriminating the same semantic label but different objects. The main contributions of TADS are as follows:

- Adaptive target-awareness deep features for object detection are not affected by complex pre-training of CNN. Thus, a few data sets can realize object detection using deep learning networks. TADT algorithm compensates for the deficiency of the pre-trained deep model being unable to consider arbitrary forms in visual detection.
- We utilized a depth-wise separable method to reduce the amount of computation associated with each frame's correlation. The speed of the algorithm has been a significant promotion.
- We used a fixed depth-wise separable convolutional kernel to avoid time waste in the iteration of back propagation.

3 Tads

3.1 Target-awareness deep tracking

The TADT algorithm introduces the target-awareness method to compute weights to express the importance of object detection's deep features. Ridge loss based on gradients is trained to obtain a proportion to distinguish deep features and ranking loss combined with the ridge component is used to represent the scale-sensitive for variation of smoke shape. The TADT algorithm includes 4 parts: pretrained CNNs, target-awareness, correlation filtering and a Siamese matching network. Figure 1 shows the framework of TADT.

VGG16 has 16 layers, including 13 convolutional layers, 5 maxpooling layers, 3 fully connected layers, an input layer and output layers. In the VGG16 model, smoke video frames are treated as input. As a result, output maps of Conv4–1 and Conv4–3, including 512 deep feature maps are treated as target-awareness model input. Target-awareness using ridge loss to determine the importance degree of 512 deep feature maps of Conv4–3. The 300 channels of deep features in 512 maps are acquired as the result of features of awareness. Table 1 give a list of all symbols.

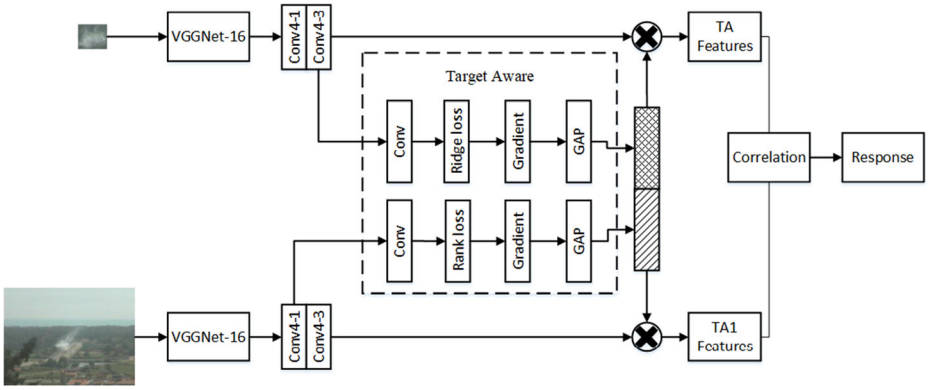

Fig. 1 The framework of TADT

Table 1 Parameters of the symbols

Serial number	Symbol	Definition
1	$Y(i,j)$	Gaussian label
2	W	The weight of regression
3	L_{reg}	The ridge loss
4	$X_o(i,j)$	The output feature maps
5	L_{rank}	Ranking loss
6	(x_i,x_j)	The scale-pairs training sample with 2 pixels stride.
7	$D_F \times D_F \times N$	Input maps
8	T_p	The number of frames with true positives
9	F_p	The number of frames with false positives.

Target-awareness uses ridge loss to research different object convolution kernels to extract characteristic information. These convolutional kernel filters provide a certain object ratio to classify object categories. In the target-awareness model, feature weights acquired by minimizing ridge loss reflect the importance of the 512 feature maps captured from the pretrained VGG16. Hence, we cannot train the VGG16 network to extract effective feature map representations for arbitrary objects in unknown scenes and avoid unnecessary bulk smoke video collection and complex network training. The ridge loss is defined as follows:

$$L_{reg} = \left\| Y(i,j) - W * X_{i,j} \right\|^2 + \lambda \|W\|^2 \tag{1}$$

Where $Y(i,j)$ is a Gaussian label defined as follows:

$$Y(i,j) = e^{-\alpha\left(\frac{c_i^2}{2\sigma_1^2} + \frac{c_j^2}{2\sigma_2^2}\right)} \tag{2}$$

Where σ is kernel width, * represents a convolution, and W is the weight of regression training to compute the contribution of feature maps. Backpropagation update weights represent the importance of feature maps. The chain rule is used to compute the derivation of L_{reg} to $X_{i,j}$ for backpropagation. The derivation using the chain rule is defined as follows:

$$\frac{\partial L_{reg}}{\partial x_{i,j}} = \sum_{i,j} \frac{\partial L_{reg}}{\partial X_o(i,j)} \times \frac{\partial X_o(i,j)}{\partial x_{i,j}} = \sum_{i,j} 2(Y(i,j) - X_o(i,j)) \times W \tag{3}$$

Where $X_o(i,j)$ is $W * X_{i,j}$ defined as output feature maps. The pretrained model extracts 512 feature maps. These feature maps are sent to a regression net to obtain the awareness feature maps characterized by the degree of importance. Each pixel gradients are acquired. According to these weights, the global gradient average pooling layer is used to obtain the instant of weights to choose 300 useful feature maps. The global gradient average pooling function is defined as follows:

$$W_i = GAP\left(\frac{\partial L_{\text{reg}}}{\partial z_i}\right) \tag{4}$$

where GAP is the global average pooling function. $\frac{\partial L_{reg}}{\partial z_i}$ is the derivation of the loss function. $L_{reg}$with respect to the i_{th} out feature map z_i. z_i is obtained by training the regression loss of the convolutional model. According to the computation of W_i, the GAP can be computed so that

we can get the importance of the out feature map. If a mapping function is constructed to demonstrate the contribution ratio of the Convolution channel, we can get an appropriate threshold to select useful channel feature response to participate in subsequent calculation in order to meet the needs of real-time response. The mapping function can be defined as follows

$$\chi_i = \varphi\left(X_{i,j}, W_i\right) \tag{5}$$

where φ is a mapping function selecting the most important channels. χ is the weight of the i-th channel.

The movement and shape of smoke depend on the wind and other environmental climates. These characteristics require the algorithm to add a scale-sensitive divisor to train the sensitive kernel filter to adapt to the scale changes. According to Li et al. [12,13], a ranking loss is proposed as follows:

$$L_{rank} = log\left(1 + \sum_{\left(x_i, x_j\right)\in\Omega} e^{f(x_i)-f\left(x_j\right)}\right) \tag{6}$$

where (x_i, x_j) is the scale-pairs training sample with 2 pixels stride. L_{rank} is minimized to adapt to the variation of the smoke shape. Ω is the set of (x_i, x_j).

In TADT, a training model is created to train the scale filter to close the extraction computation complexity for sensitive scale selection. According to the rank loss model, stochastic gradient descent (SGD) is adopted to train the rank loss to select 80 scale-sensitive deep features. The chain rule is used to compute the gradients defined as follows:

$$\begin{aligned}\frac{\partial L_{\text{rank}}}{\partial x_{i,j}} &= \frac{\partial L_{\text{rank}}}{\partial X_O(i,j)} \times \frac{\partial X_O(i,j)}{\partial x_{i,j}} \\ &= \frac{\partial L_{\text{rank}}}{\partial X_O(i,j)} \times W_{rank}\end{aligned} \tag{7}$$

where W is the convolutional kernel weight of the rank loss model. $X_o(i,j)$ is $W_{rank} \times x_{i,j}$. $\frac{\partial L_{rank}}{\partial X_o(i,j)}$ is defined as the gradient of L_{rank} relative to $f(x_{i,j})$. Scale-sensitive features are extracted according to the rank loss net. Eighty deep feature maps selected from the output maps of Conv4–1 are selected. 380 deep feature maps are finally selected by combining regression and rank loss results to represent the object characteristic and scale-sensitive expression.

3.2 Depth-wise separable convolutions

MobileNets is proposed for slight mobile embedded vision detection [8, 9, 28]. A depthwise separable strategy is built and two depthwise convolutional kernels are created to balance the latency and accuracy. MobileNet is a streamlined architecture in which a kind of factorized convolution designs depthwise separable construction. It is composed of a normal convolutional kernel called depthwise used to convolute input images, and a 1×1 convolutional kernel called pointwise applied to the output of normal convolutions. The depthwise convolution includes depthwise and pointwise. The depthwise filters the input maps, and the pointwise combines the output feature maps of the depthwise convolutions. The factorization greatly reduces the computations and decreases model complexity.

A typical convolutional layer obtains input maps defined by $D_F \times D_F \times N$ and a typical convolution kernel filter extracts output deep features defined by $F_w \times F_h \times C$. The computational consumption of typical convolutions is defined as follows:

$$W \times H \times M \times C \times F_w \times F_h \tag{8}$$

where W is the width of the typical convolutional kernel. H is the height of the typical convolutional kernel. M is the typical convolutional filter channel, and N is the number of output feature maps.

The depthwise computational cost is defined as follows:

$$W \times H \times M \times 1 \times F_w \times F_h \tag{9}$$

The pointwise computational cost is defined as follows:

$$1 \times 1 \times M \times C \times F_w \times F_h \tag{10}$$

The total computational cost of depthwise separable convolutions is defined as follows:

$$W \times H \times M \times 1 \times F_w \times F_h + 1 \times 1 \times M \times C \times F_w \times F_h \tag{11}$$

The computation reduction according to depthwise and pointwise streamline combinations can be obtained as follows:

$$\frac{W \times H \times M \times 1 \times F_w \times F_h + 1 \times 1 \times M \times C \times F_w \times F_h}{W \times H \times M \times C \times F_w \times F_h} = \frac{1}{C} + \frac{1}{W \times H} \tag{12}$$

In TADT, the cross-correlation filter [2] is applied to speed up the computations. Using the fast Fourier transform (FFT), it changes the convolutional kernel and the input feature to the frequency domain. This scheme transforms the convolution to dot product for reducing computational complexity. In this way, mathematical transformation can speed up computations. Because of the high dimensionality of multifeature maps for each frame correlation according to FFT, mathematical transformation cannot change the dimensionality of the matrix. Multifeature maps with high dimensionality require a considerable computational cost. The depthwise separable algorithm reduces the convolution computations by a streamlining operation combining two steps of depthwise and pointwise to decrease the dimensionality of the kernel. Additionally, the cross-correlation method is used to speed up the computations. This paper applies depthwise to reduce the dimensionality of the kernel to improve the efficiency of TADS effectively. Figure 2 shows the framework of TADS.

TADS uses the pretrained VGG16 model as deep features extraction model. The output maps of Conv4–1 and Conv4–3 are treated as the target aware model's inputs to extract the scale-sensitive and objectiveness deep features. The target aware model's output TA-features are composed of 300 channels of scale-sensitive deep features and 80 channels of objectiveness deep features. TA-features is exploited to construct the correlation filters by depthwise convolution. TA1-features extracted from the surveillance image are also composed of 300 channels of scale-sensitive deep features and 80 channels of objectiveness deep features. The pretrained VGG16 model extracts the deep features with target aware model. TA1-features are filtered by the filters created by TA-features to obtain the TA2-features. The filtering result TA2-features of TA-features is acquired by the correlation between the TA1-features and the depthwise convolutional kernels. The response maps are constructed by a correlation between the TA2-features and the pointwise convolutional kernels. The response maps contain three

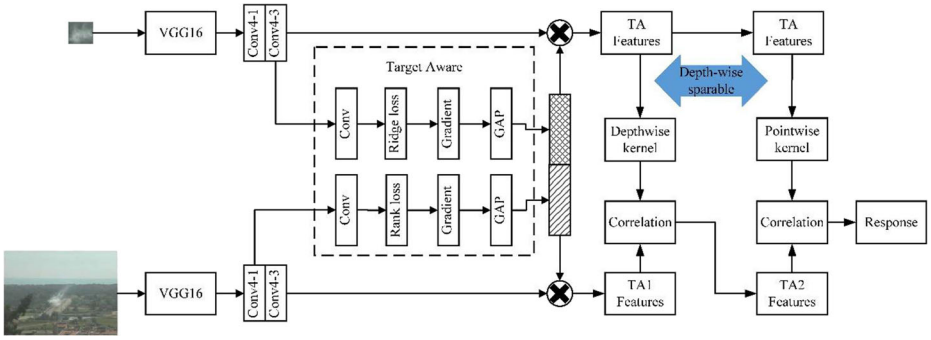

Fig. 2 The framework of TADS

response maps corresponding to the three scales of the previous target size. The response map with the max is the optimal response map. The detection results of the location and scale are determined in the optimal response map.

4 Experiments

4.1 Fire smoke video datasets

To verify the robustness and effectiveness of TADS, we select 8 fire smoke video sequences to verify the performance of the proposed algorithm TADS with the depthwise separable scheme. These data are from recognized data sets. Samples smoke videos are shown in Fig. 3.

These smoke videos are collected from web sources, and standard datasets provide by the University of Salerno. The dataset contains 17 videos with critical situations traditionally recovered as fire, such as red objects moving in the scene, smoke or clouds. To verify the effectiveness of our algorithm, we select eight videos with smokes or clouds to make the

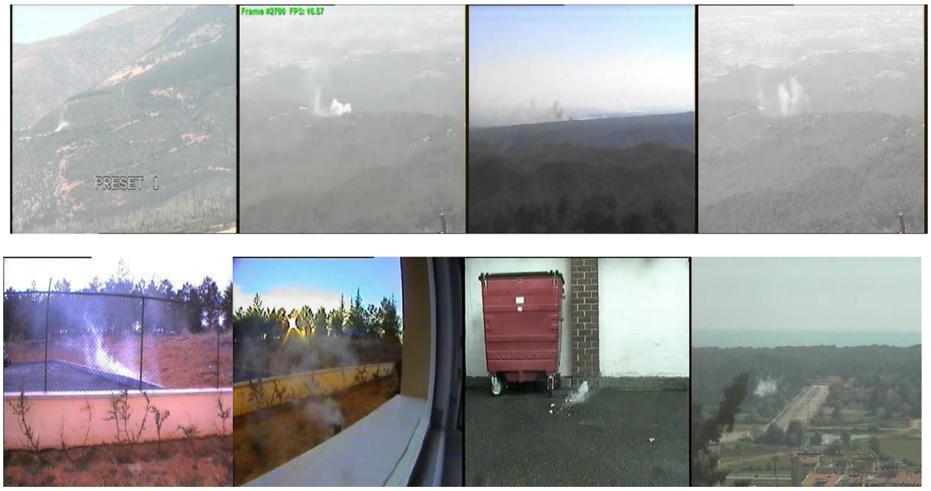

Fig. 3 Samples of the smoke videos

simulation experiment. The smoke videos of fire are selected in different conditions, including climatic conditions, scenarios, time, angle of view and resolution. The shape of the smoke is greatly affected by the wind. The robustness of detection may be affected by image resolution. Based on the above considerations, we select 8 groups of smoke video sequences to test the algorithms. Table 2 shows the summary information of the 8 groups of smoke videos. The frames define the number of images in each video, and the size defines the image resolution.

These smoke video datasets are fully considered to have similar background interference. In video 1, the fuzzy video frames are collected by the low-cost image acquisition device. In video 2, the frames include white clouds. In videos 3 and 4, the discrimination between the foreground and background of the frames is lower than those of other videos. In video 5, the shape of the smoke is affected by the wind. As for videos 6 and 8, the smoke occurs under different illumination conditions compared with other videos sequences. In video 7, the color of the wall is similar to the smoke. Video sequences of the video 1 are used to verify the ability to cope with the blurred frames of TADS. Video sequences of the video 2, 3, 4, and 7 are used to verify the ability to deal with the similarity interference. Violent movement and deformation of the smoke occurring in video 5 requires the TADS with the ability to update features effectively. Video sequences of the video 6 and 8 are used to verify the robustness of TADS under different illumination condition.

4.2 Experimental performance analysis

We use smoke videos to compute the precision of TADT and TADS. The experimental visualization results of TADT are shown in Fig. 4. These demos are chosen from the smoke videos that are selected when the algorithm runs. The visualization of TADS is shown in Fig. 5.

The experiment operation is implemented in Ubuntu 16.04 with TensorFlow on a PC with 32G memory, an Intel i7 3.7 GHz CPU and a GTX 1080 GPU. Smoke video collection and pre-processing are implemented in Win10 with MATLAB 2018a.

Figure 6 shows the precision of TADT and TADS. In Fig. 6, the precision is defined as follows:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{13}$$

where T_p is the number of frames with true positives and F_p is the number of frames with false positives. According to Fig. 6, TADS achieves the best performance in videos 1, 2, 3 and 5. The TADT achieves the best performance in videos 4, 6, 7 and 8. There is a large difference in detection accuracy between the TADT and TADS in video 5 and video 6. In video 5, the accuracy of TADT is lower than that of TADS. Because the wind affects the appearance of the smoke seriously in video5, the deep features extracted by VGG16 are different between the target and the candidates. In video 6, the accuracy of TADT is higher than the accuracy of TADS because the difference of

Table 2 Parameters of the fire smoke video datasets

	Video1	Video2	Video3	Video4	Video5	Video6	Video7	Video8
Frames	120	262	284	284	323	173	215	250
Size	325×288	320×240	720×576	720×576	352×288	320×240	320×240	320×240

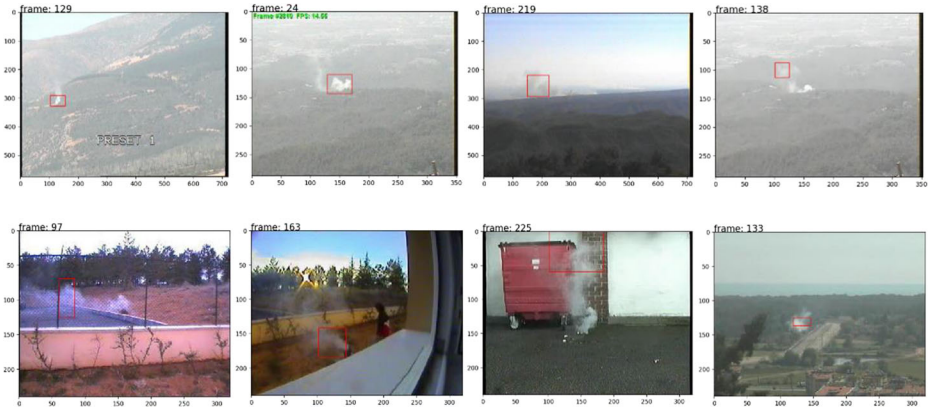

Fig. 4 Detection results visualization of TADT

the deep features between the target and the candidates is minor. Except for video 5 and video 6, the other videos' difference in detection accuracy is not apparent. The difference in the mean detection accuracy between the TADT and TADS is also not noticeable. Because the algorithms use the same feature extraction strategy, the MAP (Mean Average Precision) of TADS is the same as the TADT. Target-awareness deep features are extracted to collect robustness and semantic information. The target-awareness deep features are robust to appearance and scale changes.

The precision of the fire smoke detection acquired based on the videos are shown in Fig. 7. Figure 7 shows the MAP of the detection algorithms using the deep learning architecture and traditional pattern recognition algorithms. According to Fig. 7, the mean average precisions of TADT and TADS are higher than other algorithms, such as the 86% of YOLOv3 and 91.88% of Faster-RCNN. Figure 7 shows that the TADS performs more excellent than other deep learning algorithms and traditional smoke detection algorithms, except for TADT. Many datasets train the Faster-RCNN and the saliency detection. Figure 7 shows the higher accuracy of TADS and TADT for the blur and interference factors of videos. TADS with lower false alarm rate can be used to realize smoke detection settings in real scenes.

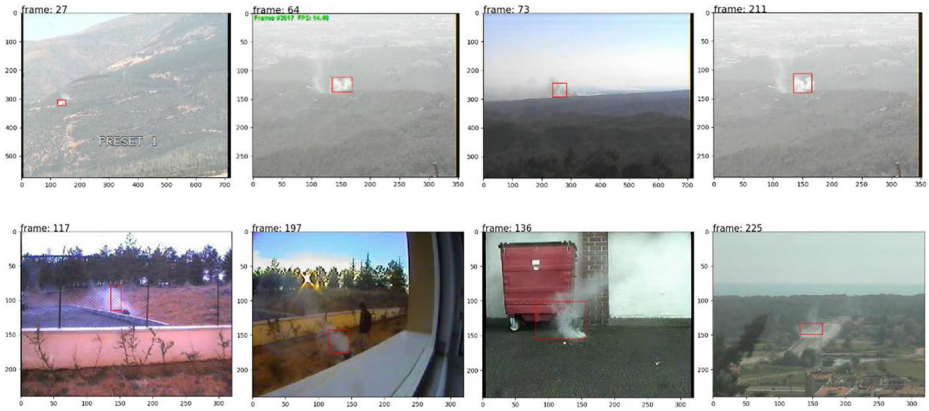

Fig. 5 Detection results visualization of TADS

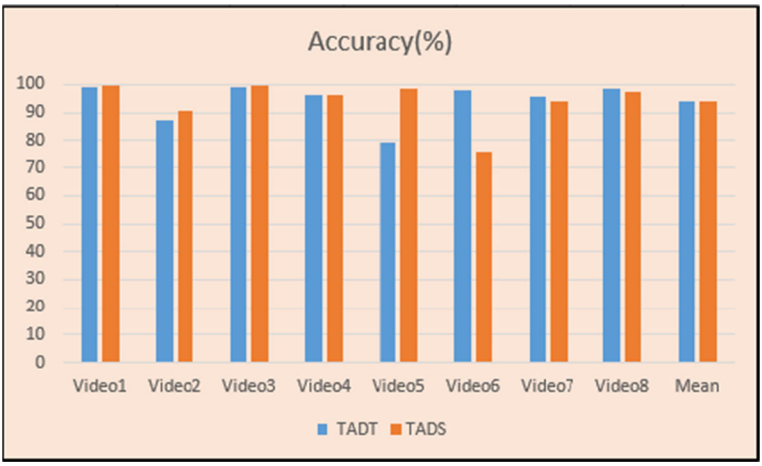

Fig. 6 Detection precision of TADT and TADS

Figure 8 shows the location error threshold-Precision curve of the TADT and the TADS based on the depthwise scheme. In Fig. 8, the curve of TADS is smoother than the TADT because we use depthwise to sharply reduce the computations, which creates more nondeterminacy for the computation of response feature maps. In Fig. 8, the location error threshold is the center Euclidean distance between the prediction bounding box and the ground truths, which are standard centers of the bounding box.

Figure 9 shows the comparison results of speed between TADS and other algorithms based on the smoke video dataset. TADS performs effectively against TADT on this dataset. According to Fig. 9, the FPS (Frames Per Second) of TADS is approximately twice that of the TADT, because TADS introduces the depthwise separable method to enhance real-time performance. The minimum frame rate of TADS can achieve approximately 86 FPS. The minimum FPS of TADS is higher than other algorithms except for TADT. The experimental results show that TADS can realize real-time smoke detection. This demonstrates the

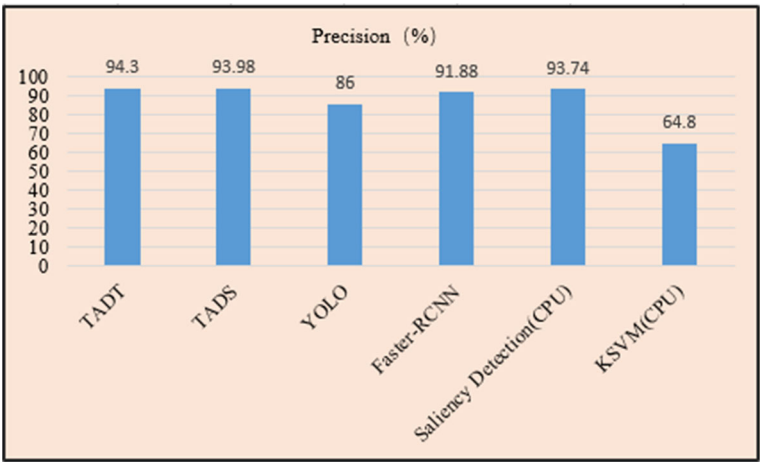

Fig. 7 The MAP precision of smoke detection algorithms

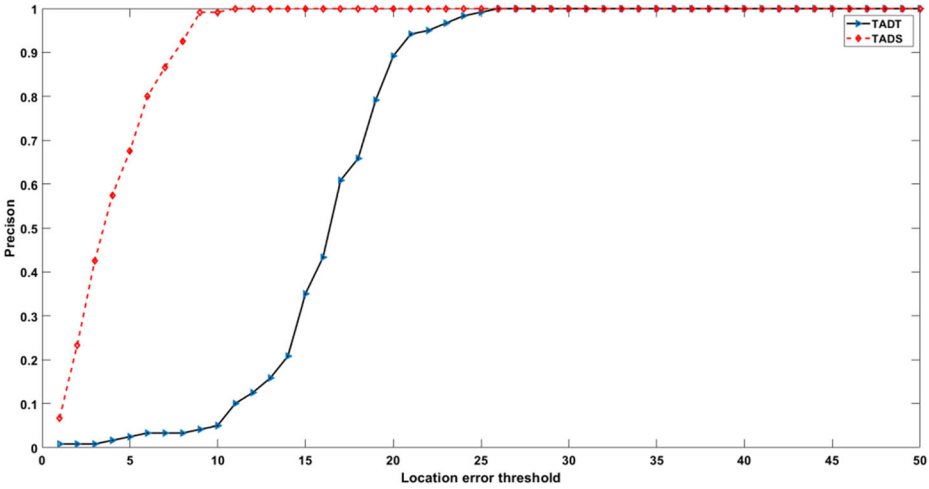

Fig. 8 The Location error threshold-Precision curve of TADT and TADS

effectiveness of TADS proposed in this paper. Overall, all experimental results demonstrate that TADS performs well in accuracy, robustness and running speed.

In this paper, we compared the accuracy and FPS of the traditional and proposed algorithms to testify that our proposed algorithm could achieve superior accuracy and speed. Figure 7 shows the MAP of the detection algorithms using deep learning architecture and traditional pattern recognition algorithms. The MAP accuracy of TADS is the same as TADT, but the speed of TADT is higher than TADT shown in Fig. 9. Figure 8 shows the location error threshold-Precision curve of the TADT and the TADS based on the depthwise scheme. In Fig. 8, the curve of TADS is smoother than the TADT because we use depthwise to sharply reduce the computations, which creates more nondeterminacy for the computation of response feature maps.

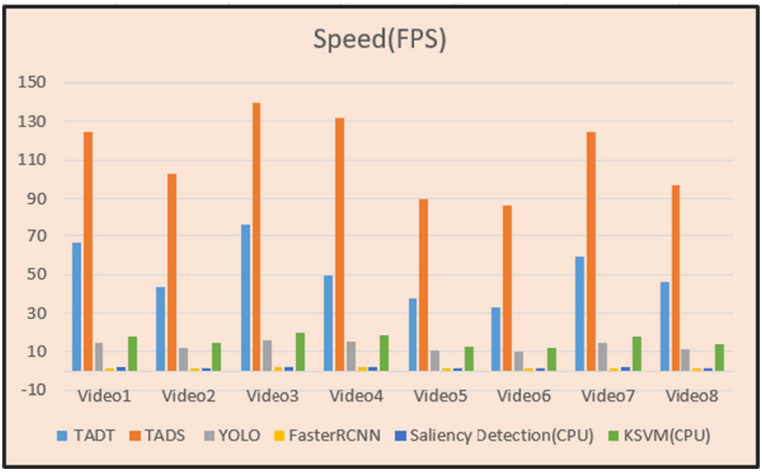

Fig. 9 The speed comparison of TADS and other algorithms

5 Conclusion

This study proposed an algorithm with a target-awareness and depthwise separable mechanism to realize fire smoke detection. The deep features extracted by the pretrained deep model included redundant information inevitably. A Target-aware scheme was exploited to delete the redundant deep features. The most useful deep features were robust to appearance and scale changes. Three hundred channels of scale-sensitive deep features and 80 channels of objectiveness deep features were selected and combined to extract the smoke and the suspected regions' features. Experimental results demonstrated that the combined deep features had strong robustness. In order to improve the efficiency of the algorithm, the depthwise mechanism was employed. The depthwise separable mechanism was composed of depthwise and pointwise convolutions to enhance real-time performance. We adopted the depthwise separable method to reduce the number of computations associated with each frame's correlation. The speed of the algorithm had been significantly improved. We used a fixed depthwise separable convolutional kernel to avoid wasted time in backpropagation iterations. The algorithm proposed in this paper could achieve the best detection accuracy of 93.98%, and the minimum 86 FPS could meet the real-time requirements. The experimental results show that our TADS algorithm has excellent performance compared with other detection algorithms. In terms of speed, it is faster than the latest detection algorithm, and its speed is comparable to that of tiny YOLO. In the accuracy comparison, the accuracy of TADS algorithm is significantly higher than that of the optimal detection algorithm, such as Faster-RCNN, which has Excellent detection performance.

TADS with the depthwise scheme increased the speed of detection comparing with TADT and other algorithms. However, the accuracy of detection was lower than other algorithms in some video sequences under different conditions. Although the depthwise scheme could decrease the number of parameters, but some useful information was delated in constructing the response map. In future work, we plan to use more output maps of convolution layers to describe the smoke to balance speed and accuracy. We also plan to use Jetson TX2 to realize this algorithm to detect fire online.

Acknowledgments We wish to thank the authors of TADT, Faster-RCNN and YOLOv3 for providing source code and the University of Salerno for providing the fire detection dataset. This research was funded by the Foundation of Henan Educational Committee grant number 16A413009 and 13B413037; the National Natural Science Foundation of China grant number 61973105 and 61573130; the Fundamental Research Funds for the Universities of Henan Province grant number NSFRF200504; the Key Technologies R\&D Program of Henan Province of China grant number 192102210073 and 212102210145.

References

1. Appana DK, Islam R, Khan SA, Kim JM (2017) A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems. Inf Sci 418:91–101
2. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2544–2550
3. Bu F, Gharajeh MS (2019) Intelligent and vision-based fire detection systems: a survey. Image Vis Comput 91:1–15

4. Chunwei T, Qi Z, Guanglu S, Zhichao S, Siyan L (2018) FFT consolidated sparse and collaborative representation for image classification[J]. Comput Eng Comput Sci 43(2):741–758
5. Gaur A, Singh A, Kumar A et al (2020) Video flame and smoke based fire detection algorithms: a literature review[J]. Fire Technol:1–38
6. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
7. Guoliang Y, Zhengwei H, Jun T (2018) Robust visual tracking via incremental subspace learning and local sparse representation[J]. Comput Eng Comput Sci 43(2):627–636
8. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861
9. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al (2019) Searching for mobilenetv3. Proceedings of the IEEE International Conference on Computer Vision, pp 1314–1324
10. Hu Y, Lu X (2018) Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. Multimed Tools Appl 77:29283–29301
11. Jia Y, Yuan J, Wang J, Fang J, Zhang Q, Zhang Y (2016) A saliency-based method for early smoke detection in video sequences. Fire Technol 52:1271–1292
12. Li Y, Song Y, Luo J (2017) Improving pairwise ranking for multi-label image classification. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3617–3625
13. Li X, Ma C, Wu B, He Z, Yang MH (2019) Target-aware deep tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1369–1378
14. Lin G, Zhang Y, Xu G, Zhang Q (1827–1847) Smoke detection on video sequences using 3D convolutional neural networks. Fire Technol 2019:55
15. Lin G, Zhang Y, Xu G, Zhang Q (2019) Smoke detection on video sequences using 3D convolutional neural networks. Fire Technol 55(5):1827–1847
16. Luo Y, Zhao L, Liu P et al (2017) Fire smoke detection algorithm based on motion characteristic and convolutional neural networks. Multimed Tools Appl 77:15075–15092
17. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
18. Mao W, Wang W, Dou Z, Li Y (2018) Fire recognition based on multi-channel convolutional neural network. Fire Technol 54:531–554
19. Morerio P, Marcenaro L, Regazzoni CS, Gera G (2012) Early fire and smoke detection based on colour features and motion analysis. 2012 19th IEEE International Conference on Image Processing. IEEE, pp 1041–1044
20. Muhammad K, Ahmad J, Mehmood I, Rho S, Baik SW (2018) Convolutional neural networks based fire detection in surveillance videos. IEEE Access 6:18174–18183
21. Pan H, Badawi D, Zhang X et al (2019) Additive neural network for forest fire detection. Signal, Image and Video Processing, pp 1–8
22. Park KM, Bae CO (2020) Smoke detection in ship engine rooms based on video images. IET Image Process 14(6):1141–1149
23. Park MJ, Ko BC (2020) Two-step real-time night-time fire detection in an urban environment using static ELASTIC-YOLOv3 and temporal fire-tube. Sensors 20(8):1–17
24. Pundir AS, Raman B (1943–1960) Deep belief network for smoke detection. Fire Technol 2017:53
25. Pundir AS, Raman B (2019) Dual deep learning model for image based smoke detection. Fire Technol 55(6):2419–2442
26. Ren S, Girshick R, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149
27. Saeed F, Paul A, Karthigaikumar P et al (2019) Convolutional neural network based early fire detection. Multimed Tools Appl:1–17
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. Proc IEEE Conf Comput Vis Pattern Recognit, pp 4510–4520
29. Sharma J, Granmo OC, Goodwin M, Fidje JT (2017) Deep convolutional neural networks for fire detection in images. International Conference on Engineering Applications of Neural Networks. Springer, pp 183–193
30. Tian H, Li W, Ogunbona PO, Wang L (2017) Detection and separation of smoke from single image frames. IEEE Trans Image Process 27:1164–1177
31. Wei Y, Zhao J, Song W, Yong W, Zhang D, Yuan Z (2015) Dynamic texture-based smoke detection using Surfacelet transform and HMT model. Fire Saf J 73:91–101

32. Xiong Z, Caballero R, Wang H, Finn AM, Lelic MA, Peng PY (2007) Video-based smoke detection: possibilities, techniques, and challenges. IFPA, fire suppression and detection research and applications a technical working conference (SUPDET), Orlando, FL
33. 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 preprint arXiv:1709.08142
34. Xu G, Zhang Y, Zhang Q, Lin G, Wang J (2017) Deep domain adaptation based video smoke detection using synthetic smoke images. Fire Saf J 93:53–59
35. Xu G, Zhang Y, Zhang Q, Lin G, Wang Z, Jia Y, Wang J (2019) Video smoke detection based on deep saliency network. Fire Saf J 105:277–285
36. Ye S, Bai Z, Chen H, Bohush R, Ablameyko S (2017) An effective algorithm to detect both smoke and flame using color and wavelet analysis. Pattern Recog Image Anal 27:131–138
37. Yin Z, Wan B, Yuan F, Xia X, Shi J (2017) A deep normalization and convolutional neural network for image smoke detection. IEEE Access 5:18429–18438
38. Yuan F, Zhang L, Xia X, Wan B, Huang Q, Li X (2019) Deep smoke segmentation. Neurocomputing 357: 248–260