 doi:10.20944/preprints202004.0027.v1

Article

Wildfire Smoke Detection based on Depthwise Separable Convolutions and Target-Awareness

Yunji Zhao [1] 0000-0001-6024-9105, Haibo Zhang [1] , Xinliang Zhang [1]* and Wei Qian

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

* Correspondence: auyjz@hpu.edu.cn;

Abstract: Since smoke usually occurs before a flame arises, fire smoke detection is especially significant for early warning systems. In this paper, a DSATA(Depthwise Separability And Target Awareness) algorithm based on depthwise separability and target awareness is proposed. Existing deep learning methods with convolutional neural networks pretrained by abundant and vast datasets are always used to realize generic object recognition tasks. In the area of smoke detection, collecting large quantities of smoke data is a challenging task 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-aware pretrained networks are used in modelling these objects of arbitrary forms to distinguish them from unpredictable and complex environments. In this paper, this scheme is introduced to deal with smoke detection. The depthwise separable method with a fixed convolution kernel replacing the training iterations can improve the speed of the algorithm 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, is superior to the state-of-the-art methods in accuracy and speed, and can also realize real-time smoke detection.

Keywords: Wildfire smoke detection; Target-aware; Depthwise separable; Fixed convolution kernel; DSATA

0. Introduction

Smoke detection is a promising method for fire alarm systems, especially in wide-open forest environments. Automatic fire detection systems play an important role in the early detection and response of unpredictable scenes[1]. Smoke video detection and analysis tasks often have difficulty obtaining ideal performance because of the multiformity of form, swing, changing smoke colour tones, environmental illumination, and low-resolution images of forest scenes. Traditional video smoke detection methods based on pattern recognition[2] and digital image processing[3] techniques depend on obtaining ample dynamic texture[4], colour features[5] [6], optical flows[7] and spatial features[8][9]. Gubbi et al.[10] adopted a pattern recognition method that manually divides the smoke video frame into 32×32 pixels to detect smoke from datasets based on wavelets[11] and support vector machines[12]. In [13], a CIELAB colour space was used to perform a smoke chromatic feature clustering method to analyse smoke colour features. In [14], histogram of oriented gradient (HOG)[15] [16] descriptors were used to extract spatial features of smoke. Xiong et al.[17] used the adaptive Gaussian mixture model (GMM) to approximate background modelling. The values that did not match background Gaussian pixels were grouped as moving blobs using connected component analysis to detect smoke. In recent years, many machine vision tasks have made great progress in the application of realistic scenarios, gaining performance across public benchmark datasets by deep learning approaches. Video smoke detection using a relatively deep network has attracted a large number of researchers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

Smoke detection methods based on deep learning adopt the mainstream deep learning framework. In [18], the normalization and convolutional neural network (DNCNN) were applied to detect smoke in smoke video. In [19], a multichannel convolutional neural network was proposed to extract deep features of fire for fire detection. Sharma [20] pretrained two convolutional neural networks (CNNs), VGG16 and Resnet50, to detect early fires. Muhammad [21] proposed a cost-effective CNN to balance complex computations and accuracy. Xu [22] applied synthetic smoke images to solve the lack of CNN training data. In [23], a background subtraction algorithm was proposed to preprocess smoke video to significantly display smoke areas, and a deep belief network was used to classify smoke.

The deep learning framework based on the proposal of interest is a class of CNN architectures combined with a region proposal method. The region-based CNN (RCNN)[24] is a CNN extension that combines selective search to detect objects. A region proposal network (RPN) is added to a typical CNN to anchor the object region of interest. Faster R-CNN[25] was proposed for pretraining VGG16 combined with RPN to classify objects and regress bounding boxes. In [26], a Faster R-CNN was adopted to crudely extract smoke areas, and a 3D-CNN was used to classify smoke video.

The deep saliency network for smoke detection is a novel method that aims to emphasize the most important object regions in video frames. In [27], salient convolutional neural networks based on pixel-level and object-level extracted smoke saliency map information were used. In [28], a saliency detection model was applied to segment a smoke region based on pixel colour and motion features. 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 the real-time requirements of smoke detection, a depthwise separable method 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. In section 2, the related works are reviewed. In section 3, DSATA is introduced. The experimental results are presented in section 4, and the conclusion is presented in section 5.

1. Related work

Smoke detection based on deep learning methods is different from traditional image processing methods. A deep learning algorithm can extract multiclass features that are not limited to one or two typical image processing features. In [29], fully convolutional networks (FCNs) were used 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 video detection methods[30][31][32] always divide each video frame into blocks and extract stable features in each block to classify smoke or nonsmoke. The highlighted performance of these methods usually relies on robust visual object forms that can obviously 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 video and high-contrast video. Existing technical conditions cannot strictly meet the requirements of video detection for large quantities of data for small sample objects. [33] proposed synthetic smoke images to meet dataset requirements. However, in visual detection, the objects of interest can be arbitrary object classes with arbitrary forms. This means that it is impossible to complete all realistic scenarios. As a result, deep feature maps for pretraining are weak in modelling these objects of arbitrary forms for distinguishing them from unpredictable and complex environments.

In this paper, according to the target-aware deep tracking (TADT) algorithm[34], DSATA is proposed with a target-aware strategy to select useful deep features for object representation. Target awareness is realized according to regression loss. In [35], the T-SNE model showed the difference between target-aware features and original features. Pretrained deep features are less effective than target-aware deep features for discriminating the same semantic label but different objects. The main contributions of DSATA are as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

- adaptive target-aware deep features for object detection do not need to require the complex pretraining of CNNs. This means that a few datasets can realize object detection using deep learning networks. The TADT algorithm compensates for the deficiency of the pretrained deep model being unable to consider arbitrary forms in visual detection.
- we adopt the depthwise separable method to reduce the number of computations associated with the correlation of each frame. The speed of the algorithm is significantly improved.
- we use a fixed depthwise separable convolutional kernel to avoid wasted time in the iteration of backpropagation.

2. DSATA

2.1. Target-aware deep tracking

The TADT algorithm introduces the target-aware method to compute weights to express the degree of importance of deep features for object detection. Ridge loss-based gradients are trained to obtain a proportion to distinguish deep features, and ranking loss combined with the ridge component is used to represent the scale-sensitive building 3 scale for variation in smoke shape. The TADT algorithm includes 4 parts: pretraining CNNs, target-aware, correlation filtering, and a Siamese matching network. Fig.1 shows the framework of TADT.

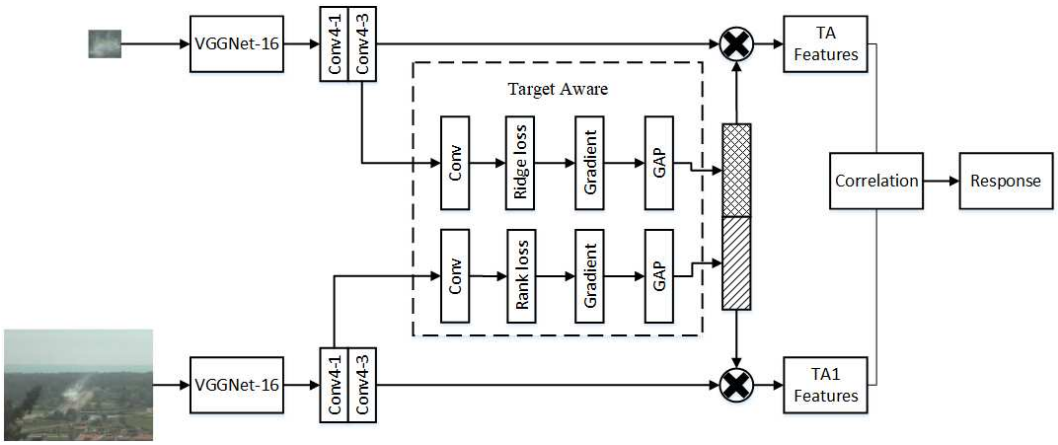

Figure 1. Framework of TADT.

VGG16 has 16 layers, which include 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, 512 deep feature maps can be acquired as target-aware model input. Target awareness uses ridge loss to distinguish the importance degree of 512 deep feature maps and filters 300 deep features in 512 maps.

Target awareness uses ridge loss to research different object convolution kernels to extract particular characteristic information. These convolutional kernel filters provide a certain object ratio to classify object categories. In the target-aware model, feature weights acquired by minimizing ridge loss reflect the importance of the 512 feature maps captured from the pretrained VGG16. This means that 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 described as follows:

$$Y(i,j) = e^{-\frac{i^2+j^2}{2\sigma^2}} \tag{2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

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, and 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:

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

where $X_o(i,j)$ is $W * x_{i,j}$ as efficiently output feature maps. The pretrained model extracts 512 feature maps. These feature maps are sent to a regression net to obtain the feature maps, which are characterized by the degree of importance. Each pixel gradient is acquired. Finally, the global gradient average pooling layer is used to obtain the instant of weights to select 300 useful feature maps according to comparison with these weights. 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. $\frac{\partial L_{\text{reg}}}{\partial z_i}$ is the derivation of the loss function L_{reg} with respect to the $i-th$ out feature map z_i obtained by training the convolutional model of regression loss.

Fire smoke shows movement and irregular shape under the influence of wind and other environmental climates. These characteristics require the algorithm to add a scale-sensitive divisor to train the sensitive kernel filter to adapt the scale changes. In [36], a ranking loss is proposed as follows:

$$L_{rank} = \log\left(1 + \sum_{\left(x_i, x_j\right) \in \Omega} e^{f\left(x_i\right) - f\left(x_j\right)}\right) \tag{5}$$

where $\left(x_i, x_j\right)$ is the scale-pairs with 2 pixel stride adjusting frames. L_{rank} loss is minimized to match the variation in smoke shape. Ω is a set of $\left(x_i, x_j\right)$.

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

$$\begin{aligned}\frac{\partial L_{\text{rank}}}{\partial x_{i,j}} &= \frac{\partial L_{\text{rank}}}{\partial X_O(i,j)} \times \frac{\partial X_O(i,j)}{\partial x_{i,j}} \\ &= \frac{\partial L_{\text{rank}}}{\partial X_O(i,j)} \times W_{rank}\end{aligned} \tag{6}$$

where W is the convolutional kernel weight of the rank loss model. $X_O\left(i,j\right)$ is $W_{rank} \times x_{i,j}$. $\frac{\partial L_{\text{rank}}}{\partial X_O(i,j)}$ is defined as the gradient of L_{rank} relative to $f\left(x_{i,j}\right)$. In this section, scale-sensitive features are extracted from the rank net of smoke, 80 deep feature maps for smoke video are selected, and 380 deep feature maps are extracted by combining regression and rank loss to represent the object characteristic and scale-sensitive expression.

2.2. Depth-wise separable convolutions

In [37][38][39], MobileNets were proposed for slight mobile embedded vision detection. A depthwise separable strategy was built, and two depthwise convolutional kernels were created to balance the latency and accuracy. MobileNet is a streamlined architecture in which depthwise separable construction is designed by a kind of factorized convolution. It is composed of a normal convolutional kernel called depthwise used to convolute input images and a 1×1 convolutional

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

kernel called pointwise applied in the output of normal convolutions. The depthwise convolution is divided into depthwise and pointwise. The depthwise filters input maps, and the pointwise combines the output feature maps of the depthwise convolutions. The factorization can greatly reduce the computations and decrease model complexity. Fig. 2 shows the typical convolution operation. The depthwise separable algorithm factorizes the kernel filter into a depthwise branch in Fig. 3 and a 1×1 pointwise branch in Fig. 4.

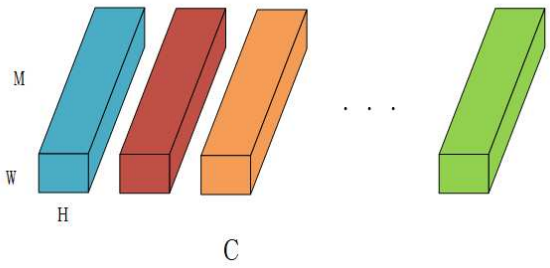

Figure 2. Typical convolution filters.

A typical convolutional layer obtains input maps as $D_F \times D_F \times N$, and a typical convolution kernel filter extracts output deep features as $F_{\mathrm{w}} \times F_{\mathrm{h}} \times C$. The computational consumption of typical convolutions is defined as follows:

$$W \times H \times M \times C \times F_w \times F_h \tag{7}$$

where W is the width of the typical convolutional kernel, H is the height of the typical convolutional kernel, M is the channel of the typical convolutional filter, and N is the number of output feature maps. Fig. 3 shows the depthwise convolution, and the pointwise convolution is shown in Fig. 4.

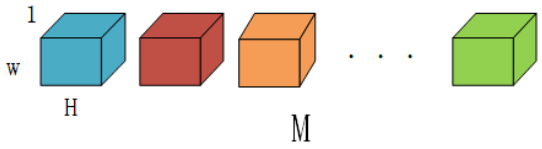

Figure 3. Depth-wise convolution filters.

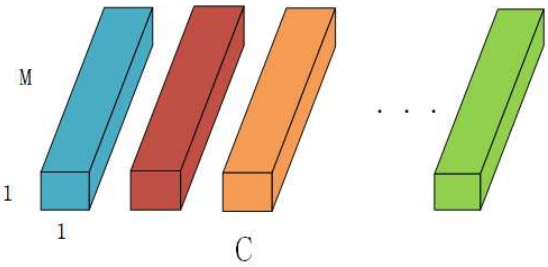

Figure 4. 1×1 point-wise convolution filters.

where the depthwise computational cost is defined as follows:

$$W \times H \times M \times 1 \times F_w \times F_h \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

The pointwise computational cost is defined as follows:

$$1 \times 1 \times M \times C \times F_w \times F_h \tag{9}$$

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

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

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

$$\begin{aligned} &\frac{W \times H \times M \times 1 \times F_w \times F_h + 1 \times 1 \times M \times C \times F_w \times F_h}{W \times H \times M \times C \times F_w \times F_h} \\ &= \frac{1}{C} + \frac{1}{W \times H} \end{aligned} \tag{11}$$

In TADT, the cross-correlation filter[40] method is applied to speed up the computations using the fast Fourier transform (FFT) to change the convolutional kernel and the input feature to the frequency domain to realize array multiplication instead of matrix operations. 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, which requires considerable computational cost. The depthwise separable algorithm reduces the convolution computations by a streamlining operation by combining two steps to decrease the dimensionality of the kernel via depthwise and pointwise operations. Additionally, the cross-correlation method is used to speed up the computations. This paper applies depthwise separability to reduce the dimensionality of the kernel to improve the architecture. Fig. 5 shows the framework of DSATA.

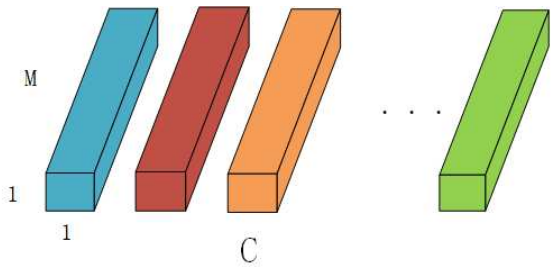

Figure 5. Framework of DSATA.

In Fig. 5, the depthwise separable algorithm is used to divide the example sample, which is the first frame smoke image in DSATA, into two steps: a depthwise convolutional kernel and a pointwise convolutional kernel. The convolutional operation first correlates the follow-up frames by extracting depthwise feature maps and then correlates the depthwise feature maps pointwise. In TADT, M×N×512 feature maps are extracted by VGG16. Three scales are added to the input smoke video. Therefore, M×N×512×3 feature maps are extracted by VGG16. The target-aware method processes the feature maps to obtain M×N×380×3 scale-sensitive and representational feature maps. In DSATA, we use average pooling to average the 380 example feature maps extracted from the target-aware strategy to M×N×1×3 feature maps as the depthwise kernel and average each feature map to 1×1×512×3 as the pointwise kernel. We process the example target-aware feature maps to segment these feature maps into depthwise filters and pointwise filters. Once the example sample is selected in TADT, it will not be changed again. Therefore, we can fix the depthwise filter and pointwise filter instead of training by deep neural networks. The fixed depthwise separable kernel pairs can also avoid the computational consumption of training of the depthwise and pointwise filters. Experimental results show that fixed depthwise separable convolutional kernel pairs can not only realize the expected

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

conclusion of approximately the same detection accuracy but also achieve a significant increase in detection speed.

3. Experiments

3.1. Fire smoke video datasets

In this section, we select 8 fire smoke video sequences to verify the performance of the proposed algorithm of the depthwise separable method of DSATA. Smoke video samples are shown in Fig. 6. These smoke videos are collected from web sources and standard datasets. Some of them are chosen in different conditions to verify the algorithm performance. The selection of smoke videos of fires considers many factors, including climatic conditions, the resolution of the camera acquisition equipment and approximate interference, coupled with the fact that smoke swings violently due to the influence of wind. The fire smoke video sequences are selected according to the above requirements, and the smoke video information is shown in Table 1.

Figure 6. Samples of smoke videos.

Table 1. Parameters of the fire smoke video datasets.

Videos	Frame Number	Size
Video1	303	352×288×3
Video2	262	320×240×3
video3	284	720×576×3
video4	284	720×576×3
video5	323	352×288×3
video6	173	320×240×3
video7	215	320×240×3
video8	250	320×240×3

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, and the influence of similar objects, such as white clouds, is added to verify the performance. In video 4, long-distance image acquisition exacerbates the degree of image blurring. At very low pixel resolution, it is still necessary to accurately detect the smoke position, which increases the experimental difficulty of the algorithm. The selection of other recognized datasets considers the effects of different experimental environments on smoke detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

3.2. Experimental performance analysis

The experiment operation is implemented in Ubuntu 16.04 with TensorFlow on a PC with 32 G memory, an Intel i7 3.7 GHz CPU, and a GTX 1080 GPU. Smoke video collection and pre-processing is implemented in Win10 with MATLAB 2018a. In this section, we use smoke videos to compute the precision of TADT and DSATA. The experimental visualization results of TADT are shown in Fig. 7. These frame demos are chosen from the smoke videos that are selected when the algorithm runs. The visualization of DSATA is shown in Fig. 8.

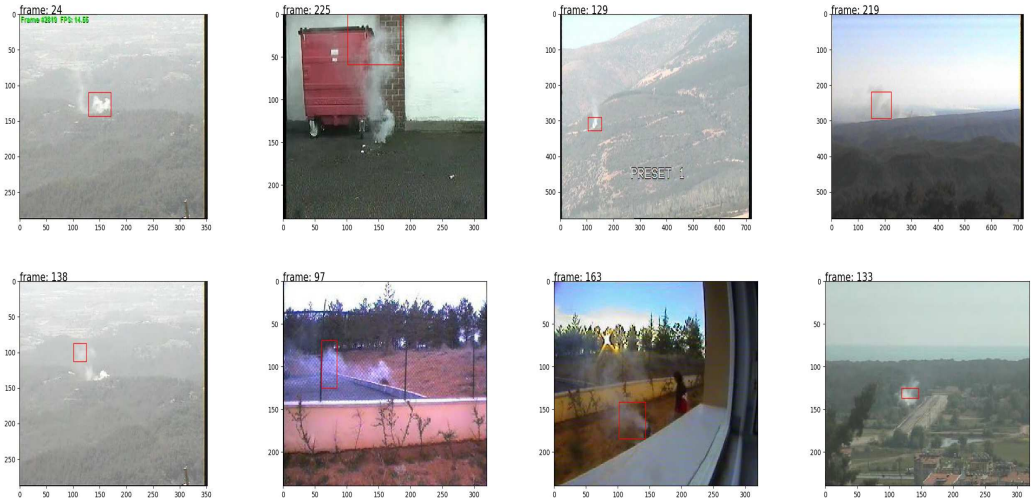

Figure 7. Detection results visualization of TADT.

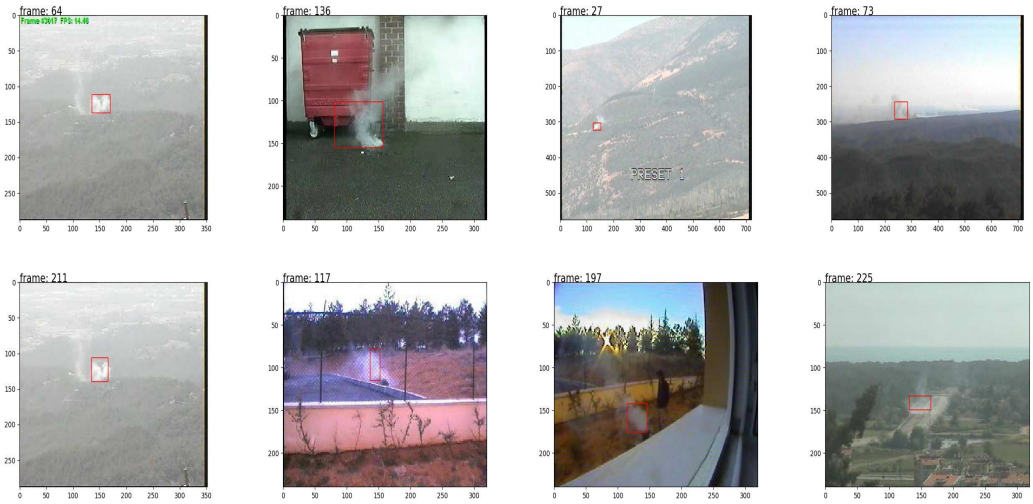

Figure 8. Detection results visualization of DSATA.

Table 2 shows the precision of TADT and DSATA. In Table 2, the precision is defined as follows:

$$Precision = \frac{Tp}{Tp + Fp} \tag{12}$$

where Tp is the number of frames with true positives and Fp is the number of frames with false positives. According to Table 2, DSATA achieves the best performance on Videos 1, 2, 3, and 5. The improved TADT used in smoke detection achieves the best performance on Videos 4, 6, 7, and 8. There is a large difference in detection accuracy between the TADT and DSATA in Video 5 and Video 6. In

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

Table 2. Detection precision of TADT and DSATA.

Videos	Precision of TADT(%)	Precision of DSATA(%)
Video1	99.17	**99.97**
Video2	87.40	**90.45**
video3	98.94	**99.65**
video4	**96.48**	96.13
video5	79.57	**98.76**
video6	**98.27**	75.72
video7	**95.8**	93.95
video8	**98.8**	97.2
Mean	**94.3**	93.98

Video 5,the accuracy of TADT is lower than that of DSATA,which maybe due to the difference both in the deep features extracted by VGG16 of example and the other frames of smokeing images.In Video 6,the accuracy of TADT is more higher than the accuracy of DSATA. The reason for this is that the distinguish deep features between example of smoke and the others is affected by the wind. Except for Video 5 and Video 6, the difference in detection accuracy in the other videos is not large. The difference in the mean detection accuracy between the TADT and DSATA is also not large. In other words, the difference in detection accuracy between the TADT and DSATA is not large because the algorithms use the same feature extraction strategy. Target-aware deep features are extracted to collect robustness and semantic information. The target-aware deep features are robust to appearance and scale changes.

Table 3 shows the precision of some smoke detection algorithms using the deep learning architecture and traditional pattern recognition. In table 3, the accuracy of DSATA is similar to TADT, but higher than the other algorithm, such as the 93.4% of DBN and the 91.88% of Faster-RCNN.etc. The data of table 3 shows that the DSATA can get excellent performance than other deep learning algorithms and traditional smoke detection algorithms. The superiority of DSATA is that DSATA can get effectively deep features without training network of deep learning. The Faster-RCNN and the Saliency Detection should be train by a large number of datasets. Table 3 shows the higher accuracy of DSATA than TADT for the blur and interference factors of video 3. According to this DSATA can be used to realize smoke detection settings in real scenes. In Fig. 9, a curve describing the TADT algorithm and the DSATA algorithm by the depthwise separable algorithm is given to show that our method can obtain better performance. In Fig. 9, the DSATA curve is smoother than the TADT curve because we use depthwise separability to sharply reduce the computations instead of performing complete computations, which may create more nondeterminacy for the computation of response feature maps. In Fig. 9, the location error threshold is the centre Euclidean distance between the prediction bounding box and the ground truths, which are standard centres of the bounding box.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2020 doi:10.20944/preprints202004.0027.v1

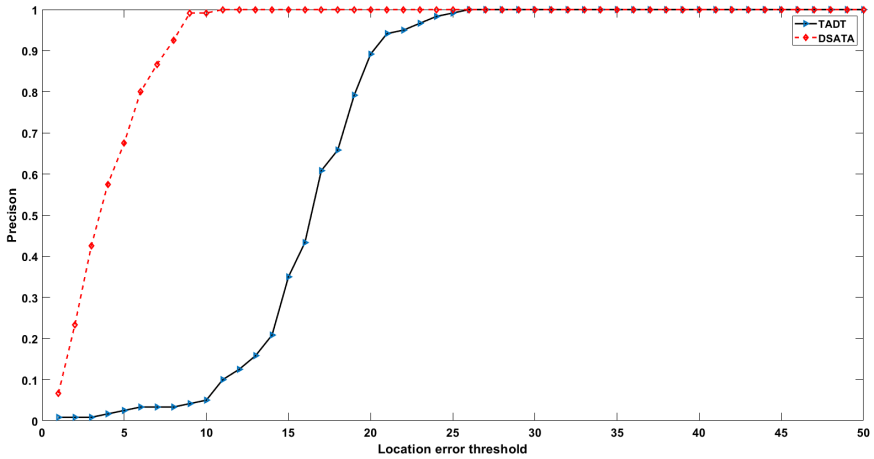

Figure 9. The Location error threshold-Precision curve of TADT and DSATA .

Table 3. Precision of smoke detection algorithms.

Algorithm	Precision(%)
TADT	**98.94**
DSATA	**99.65**
HSV+KSVM	64.8
DBN	93.4
3DCNN	93.74
Faster-RCNN	91.88
Saliency Detection	93.72

Table 4 shows a speed comparison between DSATA and TADT on the smoke video dataset. DSATA performs favourably against improved TADT on this dataset. Table 4 shows the speed of smoke detection of DSATA and TADT. According to table 4, the FPS of DSATA is approximately twice that of the TADT because DSATA introduces the depthwise separable method to enhance real-time performance. The minimum frame rate of DSATA can achieve approximately 86 FPS. The experimental results show that DSATA can realize real-time smoke detection. This demonstrates the effectiveness of DSATA proposed in this paper. Overall, all experimental results demonstrate that DSATA performs well in terms of accuracy, robustness, and running speed.

Table 4. Speed comparison of TADT and DSATA.

Videos	Speed of TADT(FPS)	Speed of DSATA(FPS)
Video1	66.414	**124.73**
Video2	43.512	**103.004**
video3	75.841	**139.666**
video4	49.679	**132.084**
video5	37.843	**89.870**
video6	32.988	**86.354**
video7	59.377	**124.770**
video8	46.327	**97.200**

4. Conclusion

In this paper, we propose an algorithm with a target-aware and depthwise separable mechanism to realize fire smoke detection. The target-aware method can extract the most useful deep features

that are robust to appearance and scale changes. The depthwise separable mechanism is composed of depthwise and pointwise convolutions to enhance real-time performance. We attempt a new method different from the mainstream methods, such as CNN, the region of proposal interest method, and saliency detection pattern recognition, which apply target-tracking algorithms to object detection. Target-aware methods can reduce the work of dataset collection, and adaptive target-aware deep features for object detection do not require the complex pretraining of CNN. We adopt the depthwise separable method to reduce the number of computations associated with the correlation of each frame. The speed of the algorithm has been significantly improved. We use a fixed depthwise separable convolutional kernel to avoid wasted time in backpropagation iterations. The experimental results show that our DSATA algorithm has excellent performance compared with other detection algorithms.

Author Contributions: Conceptualization, Yunji Zhao and Xinliang Zhang; methodology, Yunji Zhao; software, Haibo Zhang; validation, Wei Qian; formal analysis, Xinliang Zhang; investigation, Yunji Zhao; resources, Yunji Zhao; data curation, Yunji Zhao; writing–original draft preparation, Yunji Zhao; writing–review and editing, Xinliang Zhang; visualization, Haibo Zhang; supervision,Yunji Zhao; project administration, Yunji Zhao; funding acquisition,Yunji Zhao, Xinliang Zhang and Wei Qian.

Funding: 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.

Acknowledgments: We wish to thank the anonymous reviewers for their valuable suggestions and comments on this paper. We also wish to thank the authors of TADT for providing source code.

References

1. Çetin, A.E.; Dimitropoulos, K.; Gouverneur, B.; Grammalidis, N.; Günay, O.; Habiboglu, Y.H.; Töreyin, B.U.; Verstockt, S. Video fire detection–review. *Digital Signal Processing* **2013**, *23*, 1827–1843.
2. Yuan, F. A fast accumulative motion orientation model based on integral image for video smoke detection. *Pattern Recognition Letters* **2008**, *29*, 925–932.
3. Rosas-Romero, R. Remote detection of forest fires from video signals with classifiers based on K-SVD learned dictionaries. *Engineering Applications of Artificial Intelligence* **2014**, *33*, 1–11.
4. Wei, Y.; Zhao, J.; Song, W.; Yong, W.; Zhang, D.; Yuan, Z. Dynamic texture based smoke detection using Surfacelet transform and HMT model. *Fire Safety Journal* **2015**, *73*, 91–101.
5. Chen, T.H.; Wu, P.H.; Chiou, Y.C. An early fire-detection method based on image processing. 2004 International Conference on Image Processing, 2004. ICIP'04. IEEE, 2004, Vol. 3, pp. 1707–1710.
6. Chen, T.H.; Yin, Y.H.; Huang, S.F.; Ye, Y.T. The smoke detection for early fire-alarming system base on video processing. 2006 International Conference on Intelligent Information Hiding and Multimedia. IEEE, 2006, pp. 427–430.
7. Kolesov, I.; Karasev, P.; Tannenbaum, A.; Haber, E. Fire and smoke detection in video with optimal mass transport based optical flow and neural networks. 2010 IEEE International Conference on Image Processing. IEEE, 2010, pp. 761–764.
8. Hu, Y.; Lu, X. Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. *Multimedia Tools and Applications* **2018**, *77*, 29283–29301.
9. Appana, D.K.; Islam, R.; Khan, S.A.; Kim, J.M. A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems. *Information Sciences* **2017**, *418*, 91–101.
10. Gubbi, J.; Marusic, S.; Palaniswami, M. Smoke detection in video using wavelets and support vector machines. *Fire Safety Journal* **2009**, *44*, 1110–1115.
11. Ye, S.; Bai, Z.; Chen, H.; Bohush, R.; Ablameyko, S. An effective algorithm to detect both smoke and flame using color and wavelet analysis. *Pattern Recognition and Image Analysis* **2017**, *27*, 131–138.
12. Maruta, H.; Nakamura, A.; Kurokawa, F. Smoke detection in open areas with texture analysis and support vector machines. *IEEJ Transactions on Electrical and Electronic Engineering* **2012**, *7*, S59–S70.
13. Morerio, P.; Marcenaro, L.; Regazzoni, C.S.; Gera, G. Early fire and smoke detection based on colour features and motion analysis. 2012 19th IEEE International Conference on Image Processing. IEEE, 2012, pp. 1041–1044.
14. Tian, H.; Li, W.; Ogunbona, P.O.; Wang, L. Detection and separation of smoke from single image frames. *IEEE Transactions on Image Processing* **2017**, *27*, 1164–1177.

15. Ko, B.; Park, J.; Nam, J.Y. Spatiotemporal bag-of-features for early wildfire smoke detection. *Image and Vision Computing* **2013**, *31*, 786–795.
16. Park, J.; Ko, B.; Nam, J.Y.; Kwak, S. Wildfire smoke detection using spatiotemporal bag-of-features of smoke. 2013 IEEE Workshop on Applications of Computer Vision (WACV). IEEE, 2013, pp. 200–205.
17. Xiong, Z.; Caballero, R.; Wang, H.; Finn, A.M.; Lelic, M.A.; Peng, P.Y. Video-based smoke detection: possibilities, techniques, and challenges. IFPA, fire suppression and detection research and applications a technical working conference (SUPDET), Orlando, FL, 2007.
18. Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; Shi, J. A deep normalization and convolutional neural network for image smoke detection. *IEEE Access* **2017**, *5*, 18429–18438.
19. Mao, W.; Wang, W.; Dou, Z.; Li, Y. Fire recognition based on multi-channel convolutional neural network. *Fire technology* **2018**, *54*, 531–554.
20. Sharma, J.; Granmo, O.C.; Goodwin, M.; Fidje, J.T. Deep convolutional neural networks for fire detection in images. International Conference on Engineering Applications of Neural Networks. Springer, 2017, pp. 183–193.
21. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional neural networks based fire detection in surveillance videos. *IEEE Access* **2018**, *6*, 18174–18183.
22. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, J. Domain adaptation from synthesis to reality in single-model detector for video smoke detection. *arXiv preprint arXiv:1709.08142* **2017**.
23. Pundir, A.S.; Raman, B. Deep belief network for smoke detection. *Fire technology* **2017**, *53*, 1943–1960.
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
25. Ren, S.; Girshick, R.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2017**, *39*, 1137–1149.
26. Lin, G.; Zhang, Y.; Xu, G.; Zhang, Q. Smoke detection on video sequences using 3D convolutional neural networks. *Fire Technology* **2019**, *55*, 1827–1847.
27. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, Z.; Jia, Y.; Wang, J. Video smoke detection based on deep saliency network. *Fire Safety Journal* **2019**, *105*, 277–285.
28. Jia, Y.; Yuan, J.; Wang, J.; Fang, J.; Zhang, Q.; Zhang, Y. A saliency-based method for early smoke detection in video sequences. *Fire technology* **2016**, *52*, 1271–1292.
29. Yuan, F.; Zhang, L.; Xia, X.; Wan, B.; Huang, Q.; Li, X. Deep smoke segmentation. *Neurocomputing* **2019**, *357*, 248–260.
30. Calderara, S.; Piccinini, P.; Cucchiara, R. Vision based smoke detection system using image energy and color information. *Machine Vision and Applications* **2011**, *22*, 705–719.
31. Gonzalez-Gonzalez, R.; Alarcon-Aquino, V.; Rosas-Romero, R.; Starostenko, O.; Rodriguez-Asomoza, J.; Ramirez-Cortes, J. Wavelet-based smoke detection in outdoor video sequences. 2010 53rd IEEE International Midwest Symposium on Circuits and Systems. IEEE, 2010, pp. 383–387.
32. Tian, H.; Li, W.; Ogunbona, P.; Nguyen, D.T.; Zhan, C. Smoke detection in videos using non-redundant local binary pattern-based features. 2011 IEEE 13th International workshop on multimedia signal processing. IEEE, 2011, pp. 1–4.
33. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, J. Deep domain adaptation based video smoke detection using synthetic smoke images. *Fire safety journal* **2017**, *93*, 53–59.
34. Li, X.; Ma, C.; Wu, B.; He, Z.; Yang, M.H. Target-aware deep tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1369–1378.
35. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.
36. Li, Y.; Song, Y.; Luo, J. Improving pairwise ranking for multi-label image classification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3617–3625.
37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* **2017**.

38. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
39. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; others. Searching for mobilenetv3. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1314–1324.
40. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010, pp. 2544–2550.