Multimed Tools Appl (2017) 76:21651–21663
DOI 10.1007/s11042-016-4043-5

Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks

Tao Qu [1] · **Quanyuan Zhang** [2] · **Shilei Sun** [1]

Received: 30 July 2016 / Revised: 21 September 2016 / Accepted: 4 October 2016 /
Published online: 7 November 2016

Abstract In recent years, vehicle detection from aerial images obtained using unmanned aerial vehicles (UAVs) has become a research focus in image processing as remote sensing platforms on UAVs are rapidly popularised. This study proposes a detection algorithm using a deep convolutional neural network (DCNN) based on multi-scale spatial pyramid pooling (SPP). By using multi-scale SPP models to sample characteristic patterns with different sizes, feature vectors with a fixed length are generated. This avoids the stretching- or cropping-induced deformation of input images of different sizes, thus improving the detection effect. In addition, an imaging pre-processing algorithm based on maximum normed gradient (NG) with multiple thresholds is proposed. By using this algorithm, this research restores the edges of objects disturbed by clutter in the environment. Meanwhile, the raised candidate object extraction algorithm based on the maximum binarized NG entails fewer computations as it generates fewer candidate windows. Experimental results indicate that the multi-scale SPP based DCNN can better adapt to input images of different sizes to learn of the multi-scale characteristics of objects, thus further improving the detection effect.

Keywords Unmanned aerial vehicle · Vehicle detection · Multi-scale spatial pyramid · Deep convolutional neural network

1 Introduction

Unmanned aerial vehicles (UAVs) as low-cost, light-weight, imaging sensors have been constantly developed over the last decade. As a consequence, UAVs have been widely applied

✉ Shilei Sun
sunsl@whu.edu.cn

[1] International School of Software, Wuhan University, 37 Luoyu Road, Wuhan 430072, China

[2] Shanghai Spaceflight Institute of Electronic Communication Equipment, 1777 Zhongchun Road, Shanghai 201100, China

to process remote sensing data, for example, traffic monitoring, monitoring of vegetation cover [1, 6, 9, 10, 22, 26], archaeology [4, 12, 20], meteorology [21], volcano monitoring [2] and forest fire monitoring [3]. The large area remote sensing images obtained using UAVs contain a large amount of ground information. UAV-based remote sensing systems can provide complex traffic data. As a supplement to traditional traffic devices, these data have begun to be used for detecting vehicles. Vehicle detection from remote sensing images has been applied in various fields: images of road networks and distribution of vehicles in various areas can provide information for urban planning and traffic monitoring; vehicle detection and tracking from aerial images are also an important component of any video monitoring system.

In recent years, numerous algorithms have been proposed to detect and recognise vehicles from aerial images. They mainly include methods based on artificially designed feature models, shallow neural networks, and deep neural networks which include deep belief networks (DBNs) and convolutional neural networks (CNNs).

The method based on artificially designed feature models mainly refers to detecting objects with illumination invariance by using the distinguishable structures and shapes of objects. This kind of method generally presents a low recall rate (RR) and a high false alarm rate (FAR). Shallow neural networks are used on the basis of their simple characteristics with few network layers. Therefore, they present poor robustness to the displacement and rotation of objects, and high FARs in complex scenes. Deep learning algorithms contain DBNs, DCNNs, and so on. The deep neural network based vehicle detection has been proven to be the most flexible, robust, and precise method and is one of the optimal methods for vehicle detection from large area remote sensing images at present.

This research proposes a spatial pyramid pooling (SPP) based deep convolutional neural network (DCNN) for vehicle detection. When the size of images input into traditional DCNN is changed, stretching or cropping images can result in image distortion or information loss. The SPP based DCNN adopts a multi-scale spatial pyramid models for down-sampling images from characteristic patterns with different sizes, thus generating feature vectors with a fixed length. In this way, the network can directly process original images without stretching- or cropping-induced deformation, thus improving the detection effect.

2 Rapid extraction of candidate objects

Specific object detection from remote sensing images requires discovery of objects of interest in large area images and then uses complex classifiers to deal directly with original images. The large area images need to be segmented so as to reduce the size of images input into classifiers and extract the candidate windows with suspected targets. This study adopts rapid extraction for candidate objects based on binarized NG to obtain the candidate windows containing suspected targets.

2.1 Normed gradient features

The binarized NG-based rapid extraction for candidate objects generally uses generic target detection models to locate, rapidly and effectively, all objects to be detected in remote sensing images. Generally speaking, general objects have independent, favourable, closed boundaries. On this basis, the NG of images presents obvious distinctions by adjusting the windows of actual objects to a fixed size (e.g., 8×8). In this case, slight changes in closed boundaries can be presented in the characteristics of the NG. To begin with, the input images are normalised to

different quantitative sizes, followed by the calculation of the NG of adjusted images. Then, the vectors in the 8 × 8 regions of images are defined as the 64-dimentional (64D) NG of corresponding windows. Afterwards, as shown in Fig. 1, a 64D linear model is trained to select proposal windows containing targets based on the NG characteristics.

2.2 General target models for NG learning

Inspired by the fact that human visual systems can perceive objects before recognising them, 64D NG and its approximate binarisation are input into classifiers. After that, a two-stage cascaded support vector machine (SVM) is used to score the characteristics so as to acquire the model for identifying objects from image windows.

Stage 1: the weight w of classifiers is learnt using linear SVMs. To find general objects from an image, this study pre-defines and quantizes the window size (size and length-to-width ratio) so as to scan images. In addition, a linear model $w \in \mathbb{R}^{64}$ is applied to score the windows.

$$x_l = \langle w, g_l \rangle \tag{1}$$

$$l = (i, x, y) \tag{2}$$

Where, x_l, g_l, l, i and (x, y) denote the confidence score, the characteristics of the NG, position, size, and coordinates of windows, respectively; while $w \in \boldsymbol{w}$ denotes the weight of the classifier to be learned.

Stage 2: again, linear SVMs are used to learn the weights $\mathbf{v}$ and $\mathbf{t}$ of the second-level classifiers. Based on the belief scores of candidate windows acquired during Stage 1, a group of proposal windows are selected from each size i using non-maximum suppression (NMS). Therein, the windows at some sizes, for example, 512 × 512, are unlikely to contain vehicles. Therefore, the algorithm is used to screen these windows further.

$$z_l = v_i \cdot x_l + t_i \tag{3}$$

Where, $v_i, t_i \in \mathbb{R}$ and $v_i \in \boldsymbol{v}, t_i \in \boldsymbol{t}$; v_i and t_i indicate the learnt coefficient and bias term of each quantitative size i, separately; while z_l stands for the score of the window in Stage 2. Windows which have highest score are considered as input for the next step.

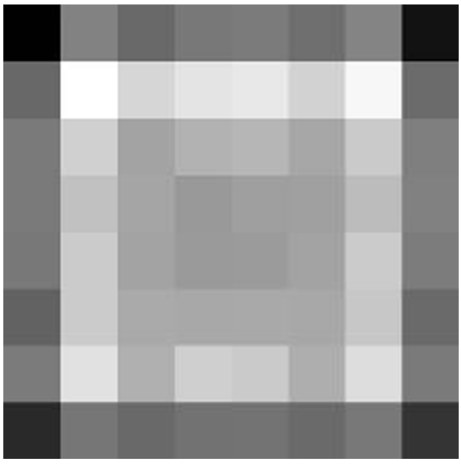


Fig. 1 NG features

3 DCNN

3.1 Pre-processing of maximum NG with multi-thresholds

As an end-to-end network structure, DCNNs can be used to process the original images directly. Nevertheless, when remote sensing images are disturbed by the environment, for instance, when objects are blocked by trees or buildings, the characteristics learnt by the neural network include noise. As a result, real information about objects is lost. To overcome this problem, a pre-processing algorithm of maximum NG with multiple thresholds is proposed.

The outline of remote sensing vehicle images contains some main information which can be used to distinguish vehicles from other objects. In this study, the unidimensional differential operator [−1, 0, 1] is used to calculate the gradient images in each channel of a colour image. The gradient with the maximum L1 norm at each pixel position is taken as the final gradient of this pixel point.

For the gradient images obtained using the maximum NG from multi-channel colour images, the edges of objects are more obvious than those in images acquired through the use of traditional algorithms. However, when objects show little clear difference from the background, or are disturbed, the edge information of the images with maximum NG is still ambiguous and we cannot effectively distinguish objects. To overcome this problem, gradient images are dealt with using a multiple threshold method so as to enhance the originally unapparent outline information. As shown in Fig. 2, no obvious difference can be found between dark cars and shadowed background and some cars are blocked by trees on the roads. After processing by multi-threshold method, the outline of dark cars is enhanced, while the textures of trees are suppressed. In this study, two thresholds (i.e., 40 and 130), are adopted. In addition, similar multiple threshold processing is also performed on grey-scale images.

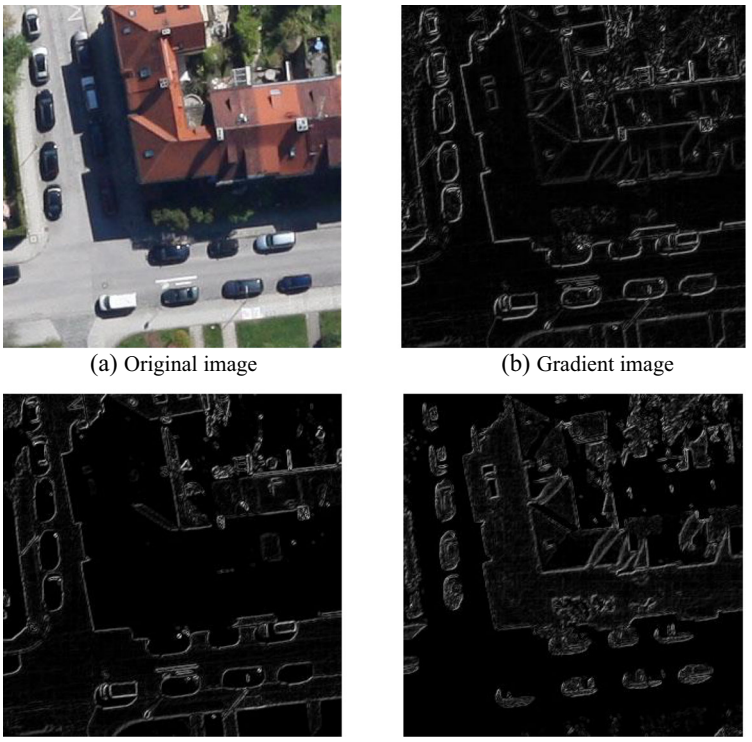
3.2 DCNN

The DCNN used in this research contains four convolution layers, four max-pooling down-sampling layers, and three fully-connected layers. Convolution layer in the DCNN contains multiple convolution kernels which output a feature map on which the weights of neurons are equal. After neurons are processed using the ReLU non-linear activation function, the output characteristics present better robustness to micro-displacement. The number of characteristic patterns in each down-sampling layer is equal to that in the former layer. In these characteristic patterns, the neurons are connected with local receptive fields of the former layer. Sampling-based dimension reduction decreases the number of neurons and extracts higher level characteristics therewith.

3.3 Convolution layers and feature mapping

For a traditional DCNN, its first four layers are convolution layers, followed by max-pooling layers: these pooling layers can also be regarded as special convolution layers. In general, they adopt sliding windows to deal with images. In addition, the last two layers are fully-connected layers and the last layer uses an N-dimensional SVM classifier to output images where N denotes the number of different categories.

Deep neural networks need to input an image of a fixed size; however, this requirement arises mainly because the fully-connected layers require input vectors of a fixed length. Besides, images of any size can be input into convolution layers as they use sliding filters to



(a) Original image (b) Gradient image

(c) Gradient image with a highthreshold (d) Gradient image with a low threshold

Fig. 2 Multiple threshold images with the maximum NG

output feature maps. These maps show a basically consistent length: width ratio to that of the input images and contain information about response intensity and spatial position.

The process of using convolution layers to generate feature maps is similar to that found using traditional methods [32] to produce characteristic images. Traditional methods adopt scale-invariant feature transform (SIFT) vectors [23] or image blocks to extract characteristics. Afterwards, these characteristics are coded through vector quantization [8, 25], sparse coding, kernel Fisher, etc. These coded characteristics contain several characteristic images and are pooled through Bag-of-Words (BoW) or using spatial pyramids. Similarly, deep convolution characteristics can also be pooled.

3.4 Multi-scale SPP layers

Also called spatial pyramid matching, SPP [5, 7, 8], as an extended version of a BoW model [11], and is one of the most successful method used in computer vision applications. It separates images into layers according to their precision and gathers local features therefrom. Before the great success of CNNs, SPP is always the key component used in competitive classification [13, 15, 17] and detection systems [18]. By combining SPP with pooling operations, this research makes it possible for neural networks to process input images of different sizes.

Images of any size can be input into the convolution layers of DCNNs to generate images with corresponding sizes. The fully-connected layers need to input vectors of fixed length,

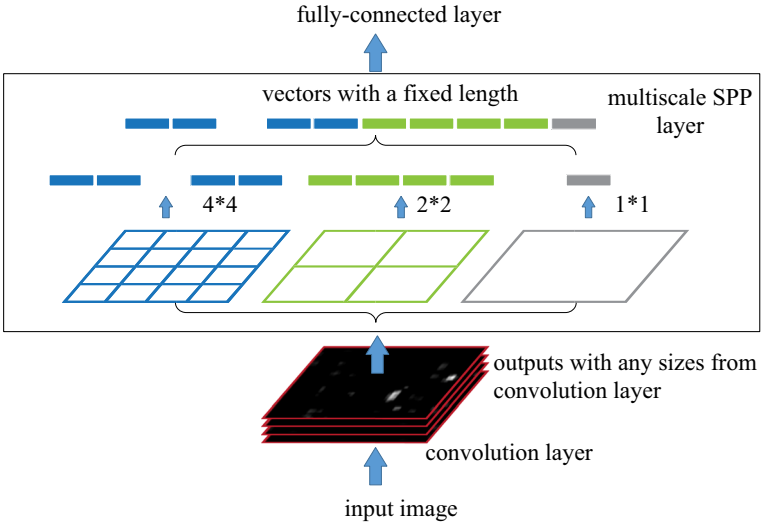


Fig. 3 The structure of a network containing SPP layers

which can be generated through pooling using the BoW method. SPP, which is improved from BoW, can be used to obtain information from local space bins through pooling. For input images of any size, the size of space bins is positively related to that of the images, while their number is fixed. This differs from the sliding windows used in traditional deep networks as the number of sliding windows depends on the size of input images.

To make deep networks adapt to input images of any sizes, multi-scale SPP layers are used to substitute max-pooling layers, as shown in Fig. 3. In each space bin, the response of filters is randomly pooled. If there are M bins, the outputs of SPP are the kM-dimensional vectors with a fixed length where k denotes the number of filters used in the last convolution layer. Vectors in fixed dimensions are the inputs of fully-connected layers.

4 Experimental results and analysis

In this study, the database used is the DLR Munich Vehicle dataset offered by the Remote Sensing Technology Institute of the German Aerospace Centre [14, 16]. The relevant images were acquired from the skies over Munich using DLR 3 K camera systems. Due to the disturbance of various factors, such as, roads, streets, trees, and similar objects, the environment in this city is far more complex than that in rural areas. Accordingly, it is challenging to use this algorithm to detect vehicles from the images in this database.

Table 1 Window number and DR based on binarized NGs

Windows number	50,000	45,000	40,000	35,000	30,000	25,000	20,000	15,000	10,000	5000
Binarized NG (%)	98.6	98.6	97.9	95.9	94.4	92.6	89.6	84.6	76.2	62.6
Sliding window(%)	16.3	15.8	15.2	13.1	12.4	11.6	10.6	9.1	8.7	7

Table 2 The RRs and FARs of vehicle detection using SPP-based DCNN

	Given Recall Rate					
	95 %	90 %	85 %	80 %	75 %	70 %
SPP-CNN	19.8	10.9	7.1	3.9	1.6	1.0
CNN	34.7	23.8	16.7	14.7	13	11.9
HOG + SVM	75.9	50.9	35.3	28	20.8	15
LBP + SVM	90.6	70.3	55.6	43.2	32.6	25.4
Adaboost	93	75.6	60.1	47	38.3	33.3

With a resolution and focal length of 5616 × 3744 and 50 mm, separately, the aerial images in the Munich Vehicle dataset are optical images obtained from 1000 m above the ground using Canon Eos 1Ds Mark III cameras installed on aircraft. These images were sampled every 13 cm on the ground and saved in JPEG format. In this database, there are total 20 images, half of which are used for training while the rest formed the test dataset. In the training images, 3418 cars and 54 trucks are contained in the positive samples with vehicles, while the test dataset includes 5928 vehicles.

To measure the performance of the DCNN algorithm, this research adopted FAR, precision rate (PR), and RR as test standards. They are defined as follows:

$$\begin{cases} \mathrm{FAR} = \dfrac{\text{number of false alarms}}{\text{number of vehicles}} \times 100\% \\ \mathrm{PR} = \dfrac{\text{number of detected vehicles}}{\text{number of detected objects}} \times 100\% \\ \mathrm{RR} = \dfrac{\text{number of detected vehicles}}{\text{number of vehicles}} \times 100\% \end{cases} \tag{4}$$

Where, the lower the FAR, the fewer objects were falsely regarded as vehicles in background windows; a higher PR implies that more vehicles are contained in the objects; while more vehicles can be detected when the RR is found to be higher. Accordingly, the algorithm aims to obtain a lower FAR and higher PR and RR as far as possible.

By counting the size distribution of vehicles in the DLR Munich Vehicle dataset, the candidate object extraction based on binarized NGs uses candidate windows at sizes of: 32, 48, 64, 80, 96, 112, and 128, separately. After being normalised, these windows measured

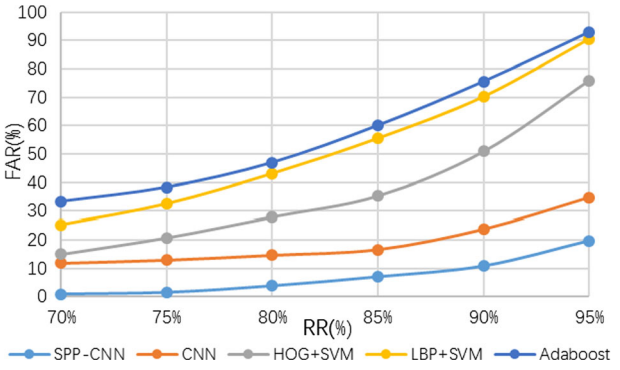


Fig. 4 The RRs and FARs of vehicle detection using SPP-based DCNN

Table 3 The RRs and PRs of vehicle detection using SPP-based DCNN

	Given Recall Rate					
	95 %	90 %	85 %	80 %	75 %	70 %
SPP-CNN	92.9	95.9	97.1	97.9	98.9	99.5
CNN	80.5	88.5	92.1	93.2	94.4	95.7
HOG + SVM	54.6	65.2	72.4	77.6	81.6	85.7
LBP + SVM	42.8	55.7	61.9	67.6	71.7	76.1
Adaboost	39.9	52.2	57.3	62.1	66.7	70.4
MVC	-	63	70	76	81	88

8 × 8. After scoring the windows by ranking SVMs in Stage 1, the approximate scores in the range of NMS spanning ±1 are calculated based on the parameters Nw = 3 and Ng = 4. As shown in Table 1, when there are 50,000 candidate windows, the binarized NG-based rapid extraction algorithm shows a DR reaching 98.6 %, while the DR of sliding window algorithm is only 16.3 %, which is far lower than that of the proposed algorithm.

In this study, the DCNN designed for vehicle detection consists of four convolution layers and four max-pooling layers in series. The images input into the network are maximum NG maps with multiple thresholds of 40 and 130, separately. In addition, the same thresholds are also found in original gradient images and multiple threshold grey-scale images. Along with the original grey-scale images, a total of six images are input to the network.

Taking the input image measuring 64 × 64 as an example, the structure of the CNNs is illustrated in detail. The convolution kernel in the first convolution layer measures 7 × 7 and the step of 1, generating 84 characteristic patterns measuring 58 × 58; max-pooling is adopted in the first pooling layer with the size of the template and step being 3 × 3 and 2, separately. With the size of the convolution kernel and step being 5 × 5 and 1, respectively, the second convolution layer produces 96 characteristic patterns measuring 24 × 24; max-pooling is used in the second pooling layer where data are processed in the same way and the layer shows the same template and step as the first pooling layer. For the convolution kernel measuring 3 × 3 and 1, separately, in the third convolution layer, 128 characteristic patterns are generated, each of which measured 10 × 10; with the size of the template and step being 2 × 2 and 1, respectively, the third pooling layer adopts max-pooling for overlapped sampling. The fourth convolution layer, which presents a 3 × 3 convolution kernel and a step of 1, separately,

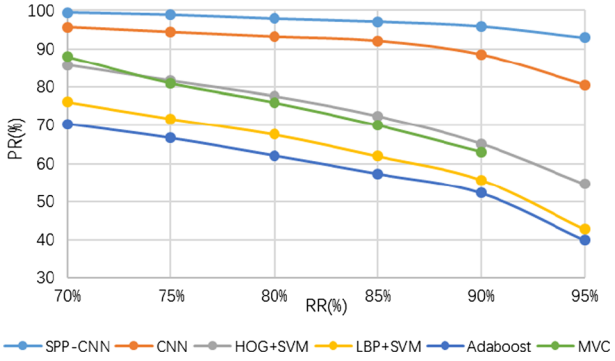


Fig. 5 The RRs and PRs of vehicle detection using SPP-based DCNN

Table 4 The RRs and FARs of vehicle detection using single-scale image

	Given Recall Rate					
	95 %	90 %	85 %	80 %	75 %	70 %
SPP-CNN	19.8	10.9	7.1	3.9	1.6	1.0
Single-scale	22.8	14.1	10	6.9	5.1	3.9

generates 128 characteristic patterns (measuring 7 × 7). While three pyramid pooling models with sizes of 1 × 1, 2 × 2, and 3 × 3, respectively, are randomly used in the fourth pooling layer to generate 128 characteristic patterns. Each characteristic pattern has a dimension of 14 and therefore there are total of 1792 dimensions. Therein, 1024 and 256 dimensions are output from the first and second fully-connected layers, respectively. During training, these two layers are learnt using a drop-out method and 2-dimensional images are output using SVM classifiers to judge whether, or not, the objects detected are vehicles. For other input images, as the structure and parameter of each network are identical, except for the size of the generated characteristic patterns, they are, therefore, not explained here.

For convenience of comparison, the detection results using six other methods based on the DLR Munich Vehicle dataset are also listed. These algorithms include: general DCNNs, histograms of oriented gradients (HOGs), SVMs, local binary patterns (LBPs), SVM [19], Adaboost, and MVC [24]. For these algorithms, grey-scale images are used as input. In the characteristic calculation using HOGs, nine orientation bins are adopted and the input grey-scale images measured 64 × 64. They can be divided into 1 × 1 + 2 × 2 + 3 × 3 + 4 × 4 + 5 × 5 = 55 blocks. Accordingly, the HOG measured 55 × 9 = 495. As for the LBPs, where $P = 8$ and $R = 1.5$ along with 58 uniform patterns and one non-uniform pattern, the feature dimension was 59 × 55 = 3245. The kernel function of SVMs is a radial basis function. Besides, five kinds of Haar characteristics and 2000 stumps are used in Adaboost. The specific detection results are illustrated in Table 2 and Fig. 4. Table 3 and Fig. 5 show the PR of the MVC algorithm [16]. Since same data to this study are adopted in MVC algorithm, Table 3 and Fig. 5 only show RR and PR results.

The experiment indicates that the detection rate of the algorithm proposed in this research is improved compared with those of traditional algorithms including: HOG + SVM, LBP + SVM, Adaboost, MVC, and general DCNNs. When the RR is given as 95 %, the multi-scale spatial pyramid network shows a detection rate of 92.9 % and a false detection rate of 19.8 %,

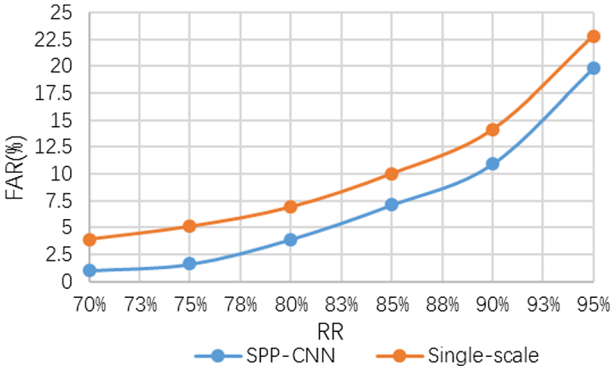


Fig. 6 The RRs and FARs of vehicle detection using single-scale image

Table 5 The RRs and FARs of vehicle detection using grey image with rotation

	Given Recall Rate					
	95 %	90 %	85 %	80 %	75 %	70 %
SPP-CNN	34.7	23.8	16.7	14.7	13	11.9
Gray image	45.5	33.1	23.8	19.2	16.9	14.3
Gray image with rotation	41.1	30.2	21.2	17.2	14.3	13.1

respectively: the detection rate and false detection rate of general DCNNs are 80.5 % and 34.7 %, separately. This is because multi-scale spatial pyramids can reduce the over-fitting problem in a network. Meanwhile, it can extract the characteristics of objects under different resolutions, thus showing a better detection effect for input images of different scales.

The impact of key parameters on the performance of the algorithm is also analysed in the experiment. Since the PR of SPP-based DCNN is already very high when RR is 95 %, we mainly use FAR as evaluation criteria. Table 4 and Fig. 6 show a false detection rate of 22.8 % when taken single-scale image as input. False alarm rate is increased owing to the characteristics extracted from single-scale image are less than multi-scale images.

The effect of gradient images with multiple threshold is tested as well. As shown in Table 5 and Fig. 7, when the RR is given as 95 % and taken grey image as input, the network shows a false alarm rate of 45.5 %. Gradient images with multiple thresholds could provide more information in certain range of grey level. Lack of such information leads to an increase of FAR. When the RR is given as 95 % and taken grey image with rotation as input, the network shows a false alarm rate of 41.1 %, lower than grey image without rotation because rotated images can help DCNN obtain more rich features, while it still presents poorer performance compared with gradient images with multiple threshold. Figure 8 shows the detection results.

5 Conclusions

This study investigates the detection of vehicles from remote sensing images in the light of the characteristics of aerial remote sensing images obtained using UAVs. By using multi-scale SPP-based DCNN to detect images of different sizes, the detection effect is improved. Although this detection algorithm has presented favourable generalisation capabilities, more universal

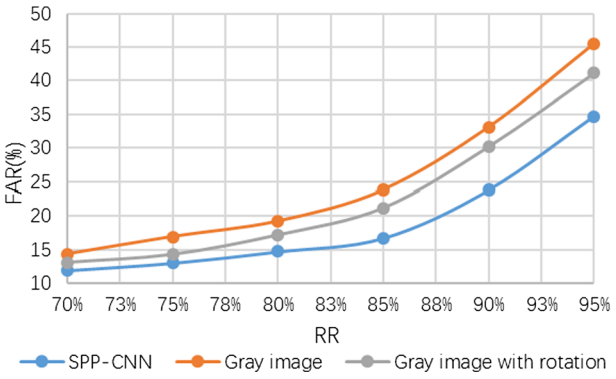


Fig. 7 The RRs and FARs of vehicle detection using grey image with rotation

Fig. 8 The detection effect of the DCNN

algorithms still need to be studied to increase the universality by reducing the number of pre-processing steps required. Besides, the background of objects in remote sensing images is generally complex as objects in city scenes are disturbed by streets, trees, architectural shadow, similar objects, etc. Therefore, it is necessary to research characteristics of stronger robustness so as to improve the detection effect of the algorithm in complex environments.

References

1. Bryson M, Reid A, Ramos F, Sukkarieh S (2010) Airborne vision-based mapping and classification of large farmland environments. Journal of Field Robotics 27(5):632–655
2. Caltabiano D, Muscato G, Orlando A, Federico C, Giudice G, Guerrieri S (2005) Architecture of a UAV for volcanic gas sampling. In: Emerging Technologies and Factory Automation, ETFA 2005. 10th IEEE Conference on, 2005. IEEE, pp 6 pp.-744
3. Casbeer DW, Kingston DB, Beard RW, McLain TW (2006) Cooperative forest fire surveillance using a team of small unmanned air vehicles. Int J Syst Sci 37(6):351–360
4. Eisenbeiss H, Zhang L (2006) Comparison of DSMs generated from mini UAV imagery and terrestrial laser scanner in a cultural heritage application. Int Arch Photogramm, Remote Sens Spat Inf Sci XXXVI-5:90e96
5. Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. In: Computer Vision, ICCV 2005. Tenth IEEE International Conference on, 2005. IEEE, pp 1458–1465
6. Grenzdörffer G, Engel A, Teichert B (2008) The photogrammetric potential of low-cost UAVs in forestry and agriculture. Int Arch Photogramm Remote Sens Spat Inf Sci 31(B3):1207–1214
7. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. Pattern Anal Mach Intell, IEEE Trans 37(9):1904–1916
8. Howard AG (2013) Some improvements on deep convolutional neural network based image classification. arXiv preprint arXiv:13125402
9. Hung C, Bryson M, Sukkarieh S (2012) Multi-class predictive template for tree crown detection. ISPRS J Photogramm Remote Sens 68(3):170–183
10. Hung C, Xu Z, Sukkarieh S (2014) Feature learning based approach for weed classification using high resolution aerial images from a digital camera mounted on a uav. Remote Sens 6(12):12037–12054
11. Krizhevsky A, Sutskever I (2012) Hinton GE Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
12. Lambers K, Eisenbeiss H, Sauerbier M, Kupferschmidt D, Gaisecker T, Sotoodeh S, Hanusch T (2007) Combining photogrammetry and laser scanning for the recording and modelling of the late intermediate period site of Pinchango alto, Palpa, Peru. J Archaeol Sci 34(10):1702–1712

13. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 2006. IEEE, pp 2169–2178
14. LeCun Y, Kavukcuoglu K (2010) Farabet C Convolutional networks and applications in vision. In: ISCAS, pp 253–256
15. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551
16. Liu K, Mattyus G (2015) Fast multiclass vehicle detection on aerial images. Geosci Remote Sens Lett, IEEE 12(9):1938–1942
17. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
18. Mathieu M, Henaff M, LeCun Y (2013) Fast training of convolutional networks through FFTs. arXiv preprint arXiv:13125851
19. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Anal Mach Intell, IEEE Trans 24(7):971–987
20. Sauerbier M, Eisenbeiss H (2010) UAVs for the documentation of archaeological excavations. Int Arch Photogramm, Remote Sens Spat Inf Sci 38(Part 5):526–531
21. Spiess T, Bange J, Buschmann M, Vörsmann P (2007) First application of the meteorological mini-UAV'M2AV'. Meteorol Z 16(2):159–169
22. Turner D, Lucieer A, Malenovský Z, King DH, Robinson SA (2014) Spatial co-registration of ultra-high resolution visible, multispectral and thermal images acquired with a micro-UAV over Antarctic Moss beds. Remote Sens 6(5):4003–4024
23. Van Gemert JC, Geusebroek J-M, Veenman CJ, Smeulders AW (2008) Kernel codebooks for scene categorization. In: Computer Vision–ECCV 2008. Springer, pp 696–709
24. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
25. Wang J, Yang J, Yu K, Lv F, Huang T (2010) Gong Y Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2010. IEEE, pp 3360–3367
26. Zhang N, Wang M, Wang N (2002) Precision agriculture—a worldwide overview. Comput Electron Agric 36(2):113–132

Tao Qu , male, born in July 1988, Ph. D candidate. He obtained bachelor degree from Wuhan University in 2010. His research interests are focused on object detection, image processing, and machine learning.

Quanyuan Zhang , male, born in October 1984, bachelor's degree. He obtained bachelor degree from Shanghai Jiao Tong University. He is employed in Shanghai Spaceflight Institute of Electronic Communication Equipment. His research interests are focused on image processing, satellites application scientific research.

Shilei Sun , male, received the Ph.D. degree in computer science from Wuhan University, Wuhan, China, in 2008. He is currently an Associate Professor in international school of software, Wuhan University. His research interests include image processing, IC design and embedded system.