doi:10.3772/j.issn.1006-6748.2019.02.004

A method for robust TV logo detection①

Pan Da(潘 达), Shi Ping②, Ying Zefeng, Hou Ming, Han Mingliang
(Information Engineering College, Communication University of China, Beijing 100024, P. R. China)

Abstract

A robust TV logo detection method based on the modified single shot multibox detector (SSD) is presented. Unlike most other existing methods which can only detect the TV logo from video frames, the proposed method can also detect the TV logo from photo pictures taken by smartphones or other smart terminals. Firstly, using a simple and effective way of collecting and labelling TV logo, a large-scale TV logo dataset used to train the detection model is built. Then, parameters and loss function of SSD are modified to make it more suitable for the task of TV logo detection. Moreover, a soft-NMS algorithm is introduced to remove the redundant overlapping boxes and obtain the final output box. And also an approach for hard example mining is designed to improve the detection accuracy. Finally, extensive comparison experiments are carried out which take into consideration different image resolutions, logo positions and environmental factors existing in real-world applications. Experimental results demonstrate that the proposed method achieve superior performances in robustness compared to other state-of-the-art methods.

Key words: single shot multibox detector (SSD), TV logo detection, TV logo dataset, loss function, hard example mining

0 Introduction

TV logo is an important symbol of TV programs and serves as a means of distinguishing TV stations and TV channels. TV logo detection plays a key role in many conventional applications such as broadcast monitoring, TV rating statistics, video archive management, video content retrieval, etc. In these applications, the TV logo detection is usually performed on key frames extracted from video sequences and thus we call it the frame-based detection.

With the popularity of smartphones, some new applications related to multi-screen interaction have emerged. For example, one may want to freely shift the programs from a TV screen to a smartphone, or one may want to find some related information with a smartphone when watching an interesting program on TV. In such scenarios, to quickly identify the TV station or TV channel is essential. One effective way is to take a photo of the TV program using a smartphone or other smart terminals and then to detect the logo with this photo picture. As a contrast, this kind of logo detection is called picture-based detection, which means the logo detection is performed on the photo pictures taken by smartphones or other smart terminals. The picture-based detection is more challenging than that of frame-based. In a photo picture, the TV logo may appear at anywhere and with any size, which makes it difficult to locate the logo position. Moreover, the environmental factors, such as the background illumination and the surrounding objects, can also be very disruptive to the logo detection.

TV logo detection is a kind of specific object detection. Related research work can be categorized into three major types: (1) low feature matching. (2) local invariant feature descriptors matching. (3) deep learning. Low features applied to representing logo basic information are usually color and shape features. Sizintsev et al.[1] matched the color histogram of the input logo image with the color histograms of templates in datasets. The Euclidean distance between two histograms is utilized to determine the similar template. Although the method is simple and easy to implement, the colors of semi-transparent TV logos change with the background, which greatly decrease the recognition accuracy. In the algorithms proposed in Refs[2,3], the contour of the TV logo is segmented from the video frame and the shape features described by Hu moment are applied to recognize the type of logo. But for hollow

① Supported by the National Natural Science Foundationof China (No. 61702466) and "Double Tops" Discipline Construction Project.
② To whom correspondence should be addressed. E-mail: shiping@cuc.edu.cn
Received on Apr. 10, 2018

logo, complex backgrounds have serious impacts on the performance of segmentation and extraction.

Local invariant feature descriptors include scale invariant feature transform (SIFT), histogram of oriented gradients (HOG), speeded-up robust features (SURF), etc. These robust local feature descriptors have been demonstrated to be invariant to displacement, deformation and occlusion. Xiao[4] proposed a SVM-based scheme which assumed that the TV logos were located in four corners of one frame and used color, edge, and key point features to train the SVM classifiers. In Ref. [5], a color-based region segmentation and candidate selection strategy was used to narrow down the candidate search space. Then the trained SVM classifiers were employed to recognize the logo type by using the HOG features. Chen et al.[6] proposed a hierarchical matching scheme in which a coarse matching based on frame differentiation was employed to narrow down the candidate space and a fine matching based on HOG was used to describe the contour features of candidate logos. Pan et al.[7] proposed a TV logo recognition scheme based on SIFT features. The original TV logo image is divided into some sub regions and the SIFT features of each sub regions are combined to represent the overall logo.

Deep learning is a branch of machine learning and has made significant achievements in object detection. Pan[8] presented a scheme which used the maximally stable extremal rgion (MSER) to generate candidate logo bounding boxes and CNN (convolutional neural network) to classify the TV logo. Unlike most other methods, this scheme dealt with the task of picture-based logo detection. Zhang[9] proposed a weakly-supervised TV logo detection system which consists of a region proposal network (RPN) and a classification network based on Fast RCNN[10]. RPN is selected as the localization network for generating numerous positive and negative candidate regions. The candidate regions along with logo class labels are used to train the fast RCNN for classification.

So far, most of TV logo detection methods are designed for frame-based applications, and their performance will drop significantly when used in picture-based scenarios. Although Ref. [8] presented a picture-based scheme, the detection accuracy is greatly affected by a large number of candidate boxes generated by MSER. And also it's not an end-to-end model because the logo locating and recognizing are implemented in two separate modules. A picture-based TV logo detection is focused and an end-to-end solution with SSD[11] as the core network is presented. As an object detector based on deep network, SSD is faster and easy to train. However, the original SSD network was designed for large and independent objects and not suitable for TV logo detection. Thus, some modifications have to be made to SSD when employed in the scheme.

The main contributions of this paper are as follows:

(1) The parameters and loss function of SSD network is modified to fit into the task of TV logo detection.

(2) The soft-NMS (soft non-maximum suppression) is introduced to remove the redundant overlapping boxes.

(3) n approach for hard example mining is designed to further improve the detection accuracy.

(4) A convenient method of collecting and labelling TV logo is developed, which is very useful for building a large-scale logo dataset.

(5) Experiments considering various image resolutions, logo positions and environmental factors are carried out to validate the robustness of our scheme.

The rest of this paper is organized as follows: Section 1 provides a detailed introduction to SSD architecture, the modifications to SSD network, soft-NMS and hard example mining. In Section 2, the experimental results and a quantitative performance assessment of the proposed method are presented. Finally, conclusions are given in Section 3.

1 Proposed TV logo detection method

1.1 Database establishment

Currently there is no public dataset on TV logo for academic research. Therefore, a large-scale dataset is set for training the detection network. However, collecting and accurately labelling data is very laborious and time-consuming, and yet a key pre-processing step in dataset building. To solve this problem, a simple and efficient collecting and labelling method is designed. Firstly, 2 000 videos of TV programs are downloaded from the Internet. These videos cover 70 classes of TV logos and three main resolutions including 1 920 $\times$ 1 080, 1 280 $\times$ 720, 720 $\times$ 480. Considering that the content between neighboring frames is extremely similar, a key frame is extracted every twenty seconds, see Fig. 1. Each key frame can be labeled automatically with only one bounding box and without any manual work based on the fact that TV logos are often located at the upper left corner. To make the detection model more robust to various environmental influencing factors, color jittering, affine transformation and rotation are taken as the data augmentation strategies. It should be noted that the rotated bounding boxes coordinates

can be computed by the rotation angles. Input images are all resized to 350 × 350 and the pixel values are normalized into the range of 0 to 1.

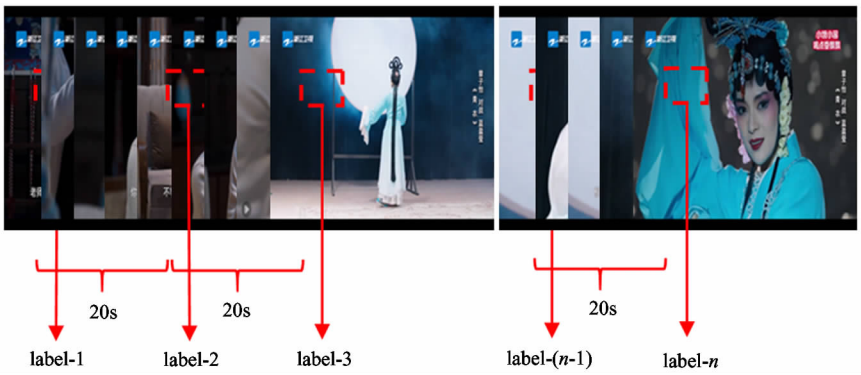


Fig. 1 The process of extracting key video frames and labelling the positions of TV logos

1.2 SSD architecture

SSD is a single-shot detector with the main idea of setting default bounding boxes on each convolution layer and training a deep network to predict the target boxes and the class probabilities of objects in those boxes. Due to its good performance on detection accuracy and speed, SSD is chosen as the core network of the scheme. Moreover, with the single-shot nature of SSD, an end-to-end model can be built for TV logo detection. The overall network architecture of SSD is shown in Fig. 2.

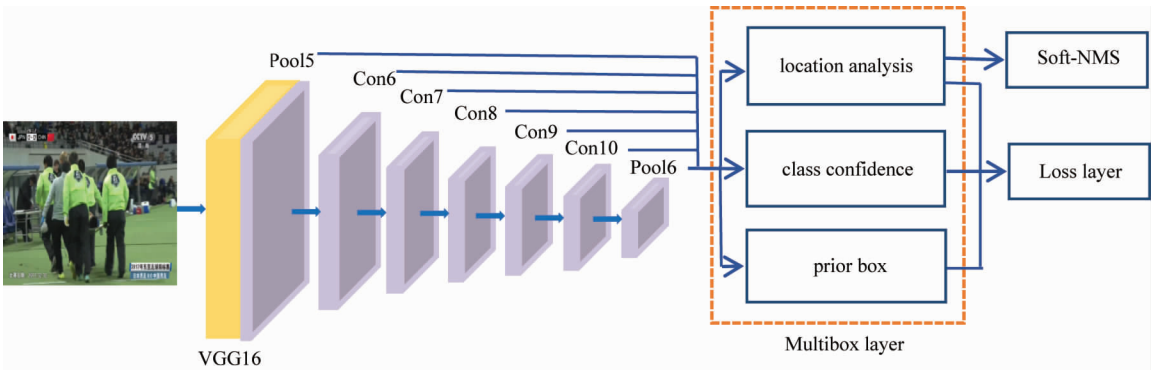


Fig. 2 SSD architecture

The base network of SSD is VGG16[12] or any other feed-forward convolutional network. Five additional convolutional layers are added after the fifth pooling layer of the base network to obtain multiscale feature maps. In this paper, to accommodate the TV logo detection, configurations of these five convolutional layers are modified as listed in Table 1. The global averaging pooling layer is selected as the last output layer. ReLU[13] is used as the activation function after each layer. Because the multiscale features from the additional layers represent different underlying information of TV logo, features from pool5, con6, con7, con8, con9, con10 and pool6 layers are combined and entered into the multibox layer which consists of three important components: location analysis component, class confidence component and prior box component. The location analysis component computes a feed-forward operation to produce a series of predicted boxes coordinates. The class confidence component selects the softmax function as the classifier for computing the confidence score of each TV logo type. The prior box component contains a series of prior boxes with different aspect-ratio in each feature map. The center coordinates (x, y) of each prior box is computed as

$$x = \frac{w + 0.5}{step\_x},\ y = \frac{h + 0.5}{step\_y},$$
$$w \in [0, w_k],\ h \in [0, h_k] \quad (1)$$

where w_k and h_k stand for the width and height of the kth layer feature map, respectively. $step\_x$ and $step\_y$ are equal to image_width/w_k and image_height/h_k, respectively.

The scale of the prior box on each feature map is

calculated as follows:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{n-1}(k-1),\ k \in [1, n] \quad (2)$$

where $s_{\max}$ and $s_{\min}$ mean the maximum and minimum ratio of prior box area to the original image area, respectively. Set $s_{\max}$ to 0.95, $s_{\min}$ to 0.2, n represents the number of feature maps.

Then the width and height for prior box on the kth layer can be computed as

$$w_\kappa^\alpha = s_\kappa \sqrt{\alpha_r},\ h_\kappa^\alpha = s_\kappa / \sqrt{\alpha_r} \quad (3)$$

where α_r is the aspect-ratio of prior boxes. Considering that most bounding boxes of TV logos are horizontal bars, the setting of aspect ratio $\alpha_r \in \{1,2,3,4,7,10\}$ is given.

Table 1 Configurations of five additional convolutional layers

Layer	Kernel number	Padding	Kernel size	Stride
Conv6	1024	1	3	2
Conv7	512	1	3	2
Conv8	256	1	3	2
Conv9	256	1	3	2
Conv10	128	1	3	2

1.3 Modified loss function

Based on observations, some TV logos are made up of two parts: the main body and the sub body as shown in Fig. 3. The main body indicates the name of TV station while the sub body gives the number of TV channel. In this paper, not only a ground truth box is used to label the whole TV logo region, but also the whole box is divided into the main body box and the sub body box. Correspondingly, the location loss function in SSD is split into two parts. The modified loss function is as follows:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc-m}(x,l,g) + \beta L_{loc-s}(x,l,g)) \quad (4)$$

$$L_{\text{conf}}(x,c) = -\sum x_{ij}^t \log c_i^t - \sum \log c_i^0,$$

$$\text{where } c_i^t = \frac{\exp(c_i^t)}{\sum_t \exp(c_i^t)} \quad (5)$$

$$L_{\text{loc}}(x,l,g) = \sum_{i \in Pos} \sum_{n \in coor} x_{ij} smooth_{L1}(l_i^n - g_j^n) \quad (6)$$

In Eq. (4), the first item $L_{conf}(x,c)$ represents the class confidence loss information, which is the softmax cross-entropy loss over multiple classes confidences (c). The second item $\alpha L_{loc-m}(x,l,g)$ represents the location loss information of the main body part, where the weight coefficient is set to 0.9. The third item $\beta L_{loc-s}(x,l,g)$ describes the location loss information of sub body part, where the weight coefficient is set to 0.75. The reason for setting different weight coefficient is that the main body part contains more critical information while the sub body part plays an auxiliary role to the total loss. The second item and the third item here are actually the Smooth L1 loss function[12] between the predicted boxes (l) and the ground truth boxed (g). In Eq. (5), $x_{ij} = 1$ denotes that the ith predicted box matches the jth ground truth box, otherwise $x_{ij} = 0$. c_i^t denotes the probability of the ith predicted box corresponding to the tth logo type. In Eq. (6), l_i^n and g_j^n represent the center point coordinate, width and height of predicted bounding boxes and ground truth boxes, respectively.

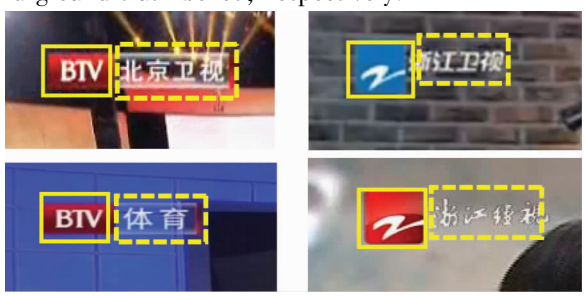


Fig. 3 The solid box refers to main body and the dashed box refers to sub body

1.4 Non-maximum suppression

For an input image, the location analysis component usually produces more than one overlapping box in TV logo region. In general, the non-maximum suppression (NMS) is used, as in SSD, to select the box with the maximum confidence score. All other boxes with a large overlap (exceeding a pre-defined threshold) with the selected box are eliminated. However, the NMS threshold setting will affect the final detection accuracy. In this paper, a novel soft-NMS[14] algorithm is used to replace the NMS algorithm. The soft-NMS uses a Gaussian function to decay the confidence scores of all other boxes according to the overlapping areas with the selected box. Hence, it needn't set the threshold and no box is eliminated in this process. Table 2 is the algorithm of soft-NMS used in the scheme.

Table 2 The pseudo code of the Soft-NMS algorithm for predicting TV logo boxes

Given:

$\boldsymbol{B} = \{b_1, b_2, b_3, \cdots, b_n\}$ is the set of predicted TV logo boxes;

$\boldsymbol{C} = \{c_1, c_2, c_3, \cdots, c_n\}$ is the set of the confidence scores of predicted TV logo boxes;

iou( ) computes the overlapping area between two boxes;

Output:

P

Table 2 Continued

Begin

$P = \{\}$

$W = \{\}$

While B is not empty **do**

$m = \text{argmax}C$

$W = b_m$

$P = P \cup W$;

$B = B - W$

for b_i in B **do**

$$s_i = s_i e^{-\frac{iou(W, b_i)^2}{\sigma}}, \ \forall \varepsilon P$$

end

end

Return P

end

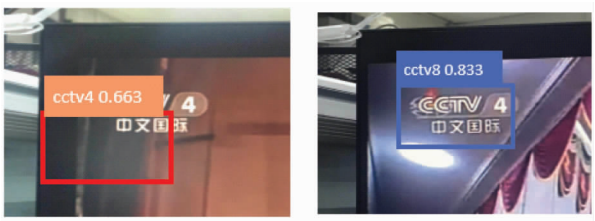
1.5 Hard examples mining

Hard examples in this paper refer to the false positive examples and false negative examples. Hard examples mining is to select the hard examples and add them to the training set to retrain the network.

For each input image, the overlapping area ratio is defined as below:

$$\text{iou} = (G \cap D)/(G \cup D) \tag{7}$$

where G denotes the area of ground truth box and D is the area of final predicted box. False positive examples refer to the instances with the overlapping areas ratio less than the threshold of 0.8. False negative examples mean that the classifier makes wrong prediction of logo types. For example, in Fig. 4(a), the box is a false positive example whose overlapping area ratio is less than the threshold. In Fig. 4(b), the logo belongs to 'CCTV4' class, but the predicted result wrongly responds to 'CCTV8' class.



(a) The false positive example (b) The false negative example

Fig. 4 The false positive and negative example

The process of hard examples mining is as follows:

Step 1 Record new TV program videos, and each video contains only one ground truth box.

Step 2 Train the detection model based on the training set of dataset until it is convergent to select model parameters with the lowest loss.

Step 3 Utilize the trained model to detect every frame of the new recorded videos. Select the hard examples by calculating the ratio of the overlapping areas and distinguishing the predicted logo classes.

Step 4 Put the selected hard examples back into the training set.

Step 5 Repeat Step 2, 3 and 4 until reaching the number of iterations.

2 Experiment results and analysis

In the experiments, the VGG16 network is selected as the base network, which is pre-trained on the ImageNet[15]. Five extra convolutional layers and a global pooling layer are added after the pool5 layer of VGG16. Hyper parameter σ of Soft-NMS is set to 0.5. The deep learning platform for training and implementation is MXNet with stochastic gradient descent, momentum 0.9, learning rate 10^{-3}, weight decay 0.0005 and batch size 32. The learning rate is decreased by a factor of 10 every 30 epochs. By extracting the key video frames, a TV logo dataset is built which has 70 $\times$2 000 $\times$30 frame samples covering 70 TV logo classes. Input images are all resized to 350 $\times$ 350. Samples in the dataset are divided into training set and validation set with the ratio of 8:2. The training process is divided into two stages: in the first stage, the parameters of base network VGG are fixed and only the extra layers' weights are trained until convergence. In the second stage, following the last epoch, all parameters of SSD are trained to fit the distribution characteristics of TV logo.

2.1 Training loss

In the training process of SSD, the cross-entropy loss of softmax and the Smooth L1 loss are used to evaluate the performance of training convergence, according to the loss function defined in Eqs(5) and (6). Fig. 5 shows the curve of the cross-entropy loss and the

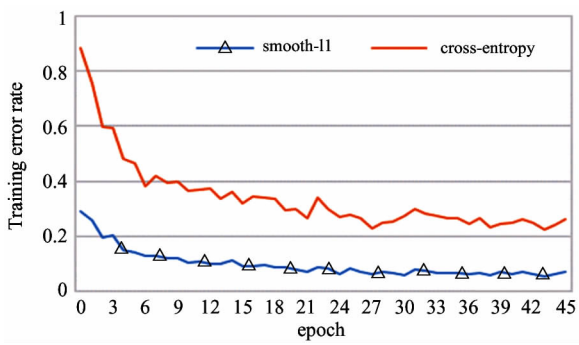


Fig. 5 Visualization of the smooth – l1 and cross-entropy loss curve with epochs.

smooth L1 loss with the change of training epochs. Fig. 6 shows the curve of the mean average precision (mAP) of the training set and the validation set with the change of training epochs. From Fig. 5 and Fig. 6, it can be seen that the model converges fast in the first 10 epochs. After the 25th epoch, losses are steady and small enough to give a very high precision in validation set, which indicates the modified loss function is appropriate for detecting TV logo.

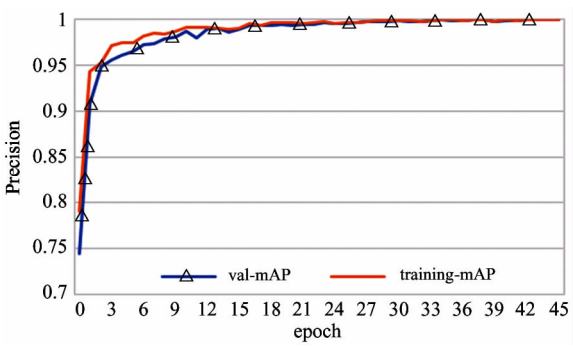


Fig. 6 Visualization of the trainingm AP and validation mAP with epochs.

2.2 Performances comparison

In this section, recall, precision, F-measure metrics are used to compare the performance of the proposed method and three other state-of-the-art methods including SIFT[7], MSER + CNN[8] and SSD[10]. The experiments are performed on the validation set of the dataset. The F-measure is calculated with recall and precision as follows:

$$F\text{-}measure = 2 \cdot recall \cdot precision / (recall + precision) \quad (8)$$

Table 3 shows the comparison results. It demonstrates the superiority of our method against other TV logo detection methods . Meanwhile , the results show that the proposed method performs much better than the original SSD in three metrics due to the modification to SSD, which makes the model more suitable for detecting TV logos.

Table 3 The results of the proposed method and other state-of-the-art methods on the TV logo dataset

Method	*Recall*	*Precision*	*F-measure*
SIFT[7]	0.773	0.813	0.792
MSER + CNN[8]	0.812	0.846	0.829
SSD[10]	0.916	0.904	0.910
The proposed method	0.938	0.915	0.926

2.3 Performances on multi-resolutions

Image resolution is a significant influencing factor in many object detection tasks. In this section, comparison experiments are performed between the method and other state-of-the-art methods on different resolutions to evaluate the precisions. Firstly, both training set and validation set are divided into three groups of samples according to the resolutions of 1 920 × 1 080, 1 280 × 720 and 720 × 480. Next, get three detection models by training the network using these three groups of training samples. Then, the testing experiments are performed on these three models using each group of validation samples. The testing results are showed in Table 4. As it can be seen from Table 4, all methods have the highest precision rate for training and testing on the same resolution. However, the performances on other resolutions decrease significantly except for SSD and the proposed method. SIFT and MSER + CNN decline by about 23% and 13%, respectively. The proposed method decreased by about 2% only, which fully proves that the proposed method has a good adaptability for different resolutions.

Table 4 The precisions of the proposed method and other state-of-the-art methods in different video resolutions

Training resolution	1 920 × 1 080			1 280 × 720			720 × 480		
Testing resolution	1 920 × 1 080	1 280 × 720	720 × 480	1 920 × 1 080	1 280 × 720	720 × 480	1 920 × 1 080	1 280 × 720	720 × 480
SIFT[7]	0.814	0.646	0.631	0.625	0.811	0.639	0.613	0.618	0.809
MSER CNN[8]	0.892	0.795	0.771	0.733	0.882	0.791	0.753	0.783	0.886
SSD[10]	0.910	0.907	0.894	0.912	0.925	0.906	0.892	0.901	0.911
The proposed method	0.931	0.913	0.902	0.918	0.937	0.914	0.899	0.905	0.925

2.4 Position independence

To validate that the proposed method is position independent for TV logo, the model on a group of images is tested in which the TV logos are located in different positions. Five volunteers are invited to randomly take pictures of TV programs using their smartphones. In the end, a total of 1 400 pictures have been collected which cover 70 classes of TV logos and each class has 20 test samples. These pictures are directly input into the network without any manual labeling. Fig. 7 shows some examples of pictures and the predicted results containing bounding boxes as well as the

corresponding logo class probabilities. From Fig. 7, the predicted results on each picture are correct no matter where the logos are located. This indicates that the proposed method is an effective position independent scheme for TV logo detection and superior to the frame-based schemes which are usually designed for dealing with the fixed position in the upper left corner of video frames.

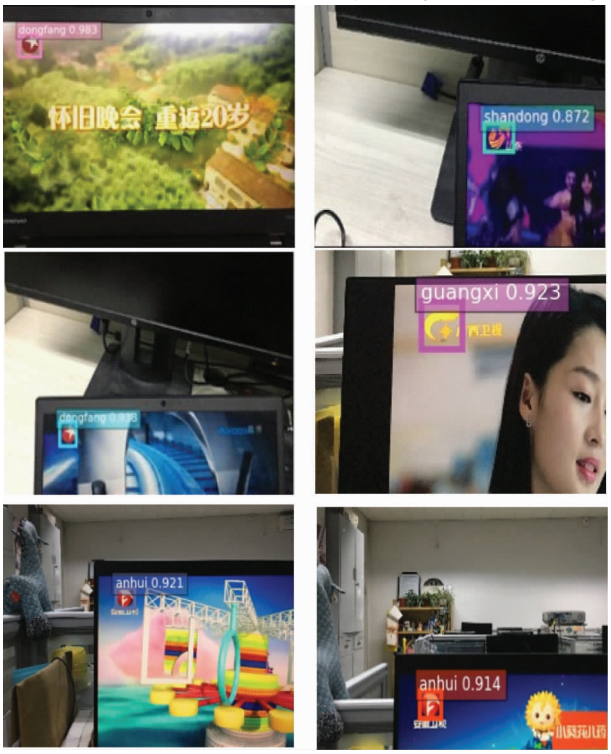


Fig. 7 The predicted results of TV logos in different positions

2.5 Experiments on environmental robustness

The environmental factors in real-world applications may bring severe interference with TV logo detection. To evaluate the environmental robustness of our method, the model on a group of images is tested which contain the main types of environmental factors: view-point, illumination, rotation and distance. Five volunteers are invited to take pictures in four different scenes designed specifically for presenting the four types of environmental factors. There are 300 test pictures in each scene and they cover most classes of TV logos. Meanwhile, the pictures in each scene are divided into three groups according to distortion degree levels of slight, bad and severe. Fig. 9 shows the pictures in three distortion levels. Fig. 9(a) – Fig. 9(c) show the view-point change. Fig. 9(d) – Fig. 9(f) show the illumination change. Fig. 9(g) – Fig. 9(i) and Fig. 9(j) – Fig. 9(l) show the rotation and distance change, respectively.

Fig. 8 shows the recall of the method and three other methods. It can be seen that the method outperforms three other methods in all four scenes. A recall of 0.93 on 'slight' level is achieved, which indicates that the model can perform well under a normal real-world application scenario. Moreover, for the 'view-point' and 'rotation', the recalls don't decrease significantly with the increasing of distortion degree owing

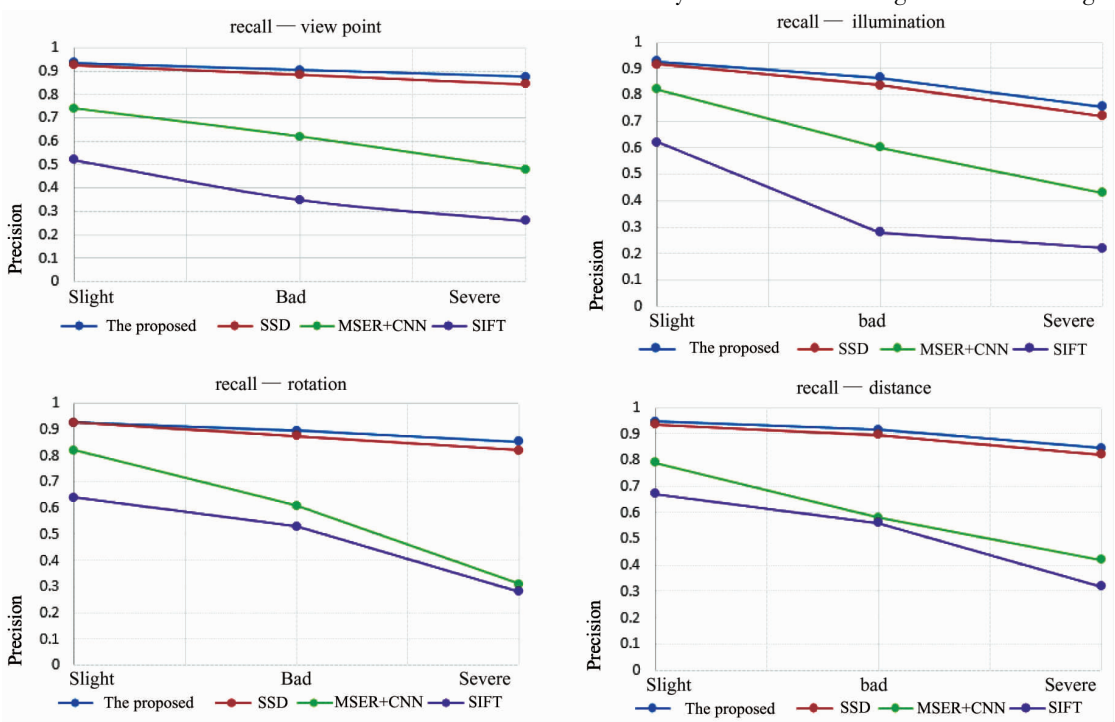


Fig. 8 The experiment results on three distortion levels

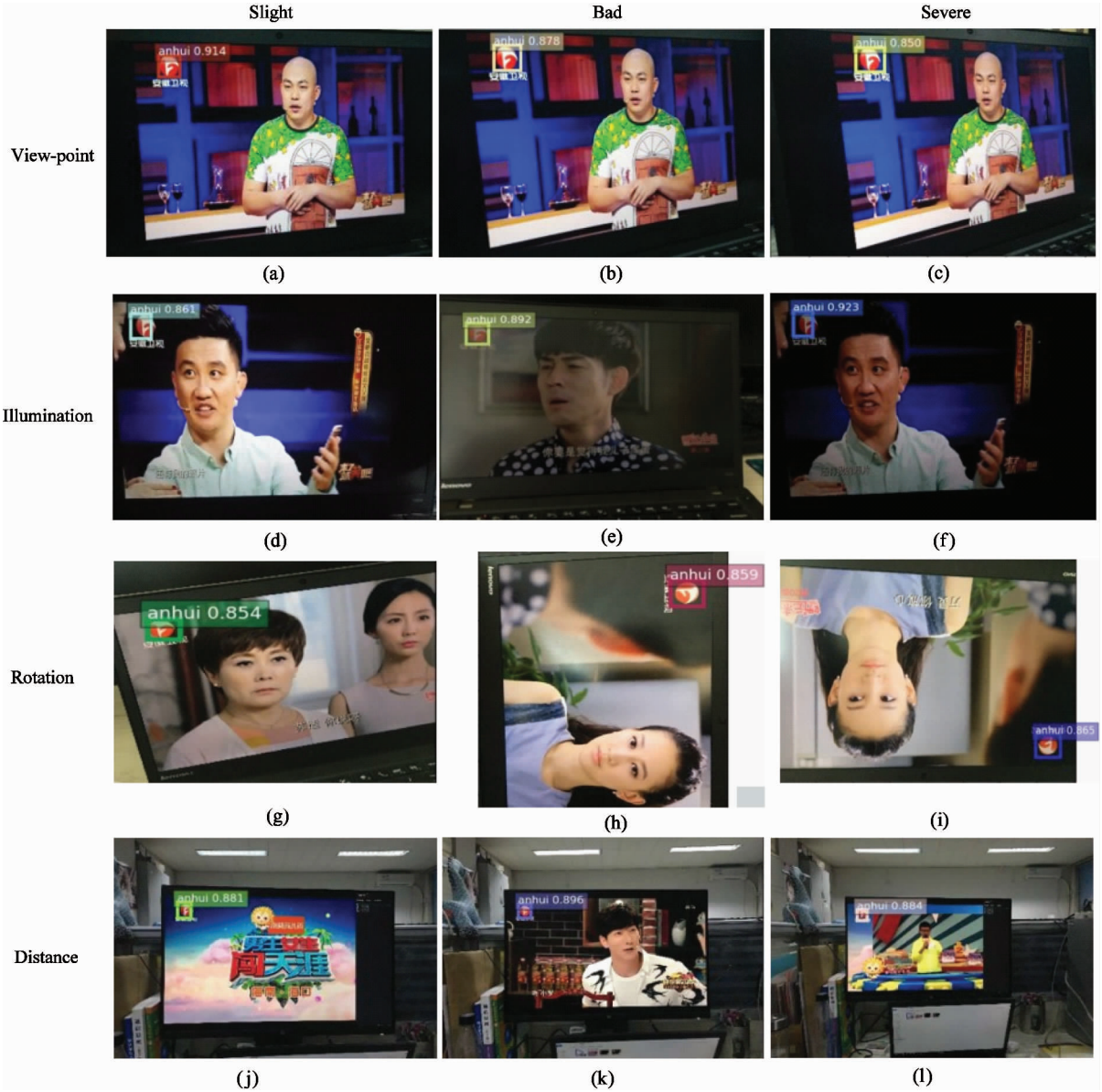


Fig. 9 The pictures on three distortion levels and in four scenes

Table 5 The comparison of the proposed method and the SSD on multi-region TV logos

TV logo names		CCTV1	CCTV2	CCTV3	CCTV4	CCTV5	CCTV6	CCTV7	CCTV8	CCTV9
SSD	*Recall*	0.914	0.917	0.881	0.928	0.933	0.907	0.922	0.911	0.893
	Precision	0.921	0.925	0.912	0.911	0.921	0.911	0.934	0.904	0.906
Our method	*Recall*	0.943	0.933	0.926	0.928	0.933	0.928	0.926	0.921	0.923
	Precision	0.933	0.932	0.932	0.911	0.930	0.915	0.934	0.915	0.936

to the data augmentation strategies of affine transformation and rotation. SSD shows almost the same performance as ours on 'slight' level, but it becomes worse on 'bad' and 'severe' level. Compared with the method and SSD, performance of MSER + CNN is much poor because MSER cannot capture TV logo areas accurately in bad environments. The method of SIFT shows the worst performance even on 'slight' level. This is because the SIFT features are difficult to be extracted from small size of TV logos. And more importantly, it is not designed for the picture-based logo detection. As a conclusion of the experiments, the proposed method has a good performance in environmental robustness.

2.6 Comparison on TV logos with multi regions

In this paper, the original loss function in SSD is modified to respond to the multi-region characteristic of

TV logos. To validate the effectiveness of the modification, the model and the original SSD model are tested on a group of video frames with multi-region logos. In general, if a TV station broadcasts its programs by multiple parallel channels, the TV logo on each channel will be made up of multi regions: the main body and the sub body. Fig. 10 shows three examples: CCTV 1, CCTV 3 and CCTV 6. Table 5 shows the recall and precision of the proposed method and original SSD on the images from nine channels of CCTV, which proves that the modified loss function performs better and is more suitable for detecting TV logos with multi regions.

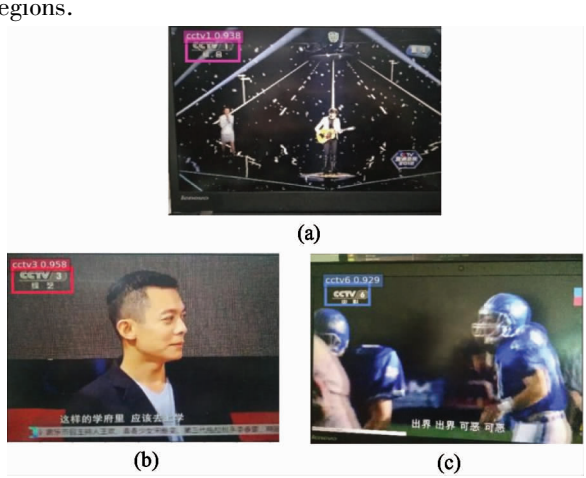


Fig. 10 Examples of multi – region TV logos

3 Conclusions

Compared with the frame-based TV logo detection, the picture-based detection is a more demanding task and thus requires a more robust detection model. In this paper, a TV logo detection method is proposed which uses SSD as the core detection network. In order to build a large-scale TV logo dataset for network training, a simple and effective way of collecting and labelling TV logo is developed. To make SSD suitable for TV logo detection, parameters and loss function of SSD are modified. In addition, the soft-NMS algorithm is employed to remove the redundant overlapping boxes and an approach for hard example mining is designed to improve the detection accuracy. To confirm the effectiveness of the proposed method, a series of experiments are designed and conducted. The results reveal that the proposed method has a superior robustness on different image resolutions, logo positions and environmental factors existing in real-world applications.

References

[1] Mikhail S, Derpanis K G, Hogue A. Histogram-based search: a comparative study[C]. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA. 2008. 1-8

[2] Zhou X Z, Shi Y C, Wang T. TV symbol recognition based on weighted Hu invariant moments in HVS color space[J]. *Journal of Nanjing University of Science and Technology*, 2005, 29(3):363-367

[3] Wang J Q, Duan L, Li Z, et al. A robust method for TV logo tracking in video streams[C]. In: IEEE International Conference on Multimedia and Expo, Toronto, Canada, 2006. 1041-1044

[4] Xiao G R, Dong Y, Liu Z X, et al. Supervised TV logo detection based on SVMS[C]. In: IEEE International Conference on Network Infrastructure and Digital Content, Beijing, China, 2010. 174-178

[5] Ye F, Zhang C, Zhang Y, et al. Real-time TV logo detection based on color and HOG features[C]. In: International Symposium on Broadband Multimedia Systems and Broadcasting, London, UK, 2013. 1-5

[6] Chen W J, Lan S Z, Xu P. Multiple feature fusion via hierarchical matching for TV logo recognition[C]. In: International Congress on Image and Signal Processing, Shenyang, China, 2015. 659-663

[7] Pan D, Shi P. A method of TV Logo recognition based on SIFT[C]. In: International Conference on Model Transformation, Guangzhou, China, 2013. 1571-1579

[8] Pan D, Shi P, He Q Z, et al. TV logo classification based on convolutional neural network[C]. In: International Conference on Information and Automation, Ningbo, China, 2016. 1793-1796

[9] Zhang Y Y, Cao X C, Wu D, et al. Weakly-supervised TV logo detection[C]. In: 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation, Hefei, China, 2017. 1031-1036

[10] Ren S Q, He K M, Ross B, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(6):1137-1149

[11] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. In: European Conference on Computer Vision, Amsterdam, Netherlands, 2016. 21-37

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv*:1409.1556V6, 2014

[13] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010. 807-814

[14] Bodla N, Singh B, Chellappa R, et al. Soft-NMS-improving object detection with one line of code[J]. *arXiv*: 1704.04503v2, 2017

[15] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. In: IEEE Conference on Computer Vision & Pattern Recognition, Miami, USA, 2009. 248-255

Pan Da, born in 1989. He is a Ph. D candidate in Communication University of China in 2018. He also received his B. S. and M. S. degrees from Communication University of China in 2012 and 2015 respectively. His research interests include the design of algorithms for image processing, computer vision and deep learning.