Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2018 doi:10.20944/preprints201801.0160.v1

Article

A lightweight video detection framework based on information theory and machine learning

Qi Liu[1], Guangtai Ding[2]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China; lq1326@i.shu.edu.cn
[2] School of Computer Engineering and Science, Shanghai University, Shanghai, China; gtding@i.shu.edu.cn
* Correspondence: gtding@shu.edu.cn

Received: date; Accepted: date; Published: date

Abstract: In recent years, many algorithms based on end-to-end deep networks have been proposed to deal with the target detection problem of videos. However, the deep network models usually consume a lot of computing resources during the procedure of analysis of videos with complex dynamic backgrounds. In this paper, a new method of object detection based on information theory is presented. Firstly, each frame in a video is converted into an effective information map by using the Harris corner detection method. Secondly, the sensitive areas in the frame are extracted by using the context information and the effective information maps of the consecutive video frames. The sensitive areas in the video frame are the candidate areas where the target objects would be appeared at high probabilities. Thirdly, the information entropy features of each sensitive area are extracted to form the feature matrix, based on which, an SVM model is trained for selecting the target areas from the sensitive areas. Finally, the locations of the objects are detected based on the target areas in the video with a complex dynamic background. As a lightweight video detection framework, the method presented in this paper can save a lot of computing resources. Experimental results show that this method can achieve good results in the benchmark of CDnet 2014.

Keywords: target detection; dynamic background; information theory; feature matrix; computing resources

1. Introduction

In the last decade, the video processing technology of target detection and tracking has made great progress. Without considering the computational cost of hardware, the most common and effective methods usually depend on the deep neural networks. The end-to-end convolutional models often have good performance in speed and accuracy [1]. Actually, the models are mainly designed for the target detection of still images. In the video processing, these models need to be optimized with the contextual information. When some interfering factors, such as illumination change, motion blur or video jitter, appear in the video, the accuracy of recognition of them would decline sharply.

The speed of video processing rests with the positioning time of suspected target areas in the video frame [2]. The computational cost of video processing depends on the complexity of the deep network model [3]. According to the regularity of the background change and the finiteness of the target category, the video processing model can be designed and modified specially for some certain scenes. For example, the NoScope model is made up of difference detector, specialized model and reference neural network [4]. Due to the special structure and the search method of the model, it can both improve processing speed and reduce operation cost. But after the complex training, the NoScope model is highly dependent on specific scene and target.

The video processing requires avoiding the use of complex deep neural network to reduce hardware cost. At the same time, it also requires exploiting the data redundancy between consecutive frames to reduce software (computation) costs. Zhu [5] uses sparse feature propagation to save these

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

cost. These propagated features are only calculated on the key frames in the video. But this kind of algorithm still needs to be improved because there is not a proper method to dynamically select the key frames from all video frames.

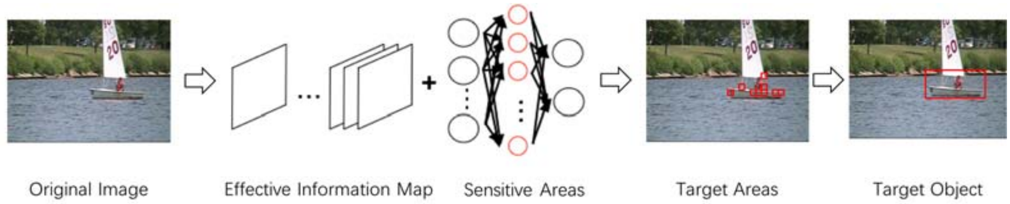


Figure 1. The flow chart of method.

In this paper, a method based on information theory is proposed to detect the video object at low cost. The flow chart of the method is shown in Figure 1. In a video with complex and changeable background, it is very difficult to distinguish the target from the noise only by the motion trajectory. But the change of local information entropy between consecutive frames is effective to identify the object. Based on these features, it is possible to design a relatively simple model for noise reduction. Thus, the hardware which supports this model to run can save a lot of cost. In addition, the sensitive areas in the video frame are more likely to contain foreground targets. So it can avoid a lot of time-consuming operations by doing intensive calculation for sensitive areas rather than that for the whole frame image. The objective of experiments in this paper is to detect objects in the video with dynamic background. And the definition of target detection in this paper is only to mark object positions, not to classify the objects. Comprehensive experiments show that the method based on information theory achieves high accuracy and significant performance.

2. Previous Work

In the past period, there has been a lot of discussion and research on target detection and object classification in the field of video processing. On the basis of the image processing algorithms, a large number of video processing algorithms has been proposed and improved. In practice, many experiments show that combining traditional feature extraction methods (like SIFT [6] and HOG [7]) and machine learning methods can solve the problem of object detection and classification well in the field of image processing [8,9]. In DPM model, the object is divided into multiple parts for extracting HOG features separately [10]. These HOG features are proved to be useful and efficient in pedestrian detection. Due to the lower computational cost and higher detection accuracy, deep neural network becomes an important method of computer vision processing in recent years. Current evidence suggests that the deep learning algorithm has an ability of approximate human beings in the field of image classification and object detection [11,12].

When all of the frames in the video are processed in the same way, the applications of deep network in image processing can be directly converted to algorithms for solving video processing tasks. Without considering the cost and the speed, the task of object recognition and segmentation in video processing can be solved well by improving the deep network algorithm (like CNN [13]) in image processing [14,15]. But as the requirements for speed and accuracy are gradually improved, more models are specially designed as end-to-end deep network models [16,17]. These models can avoid many problems about parameter optimization during the training process. Meanwhile they can also speed up by directly outputting the category and the location of object.

The end-to-end network model is considered as the key point of improving the speed of recognition and positioning. In the object recognition task of computer vision, the YOLO model abandons the thought of regional pre-processing. And it truly realizes the application of end-to-end network model [18,19]. Now, without considering the cost of hardware, the target object in video processing task can be identified and tracked in real time, such as SSD model [20]. Zhu [5] proposes the idea of transferring the convolutional feature map of sparse key frame to other video frames. This idea can take advantage of the context information well in the video stream and accelerate the operation process of the whole model by flow calculation. This method is specially designed for solving video processing tasks.

The use of end-to-end deep network needs to consume a lot of costs, such as manual marking cost and hardware cost. So before the end-to-end deep learning network becomes popular, R-CNN (regional convolutional neural network) is the primary method to solve the object recognition problem in the field of video processing. Compared to a complex deep network, R-CNN may be run with lower hardware cost but will be followed by the inefficiency of computation. The method proposed in this paper tries to solve this contradiction. It improves the speed of video processing by using sensitive area screening. At the same time, it saves hardware cost by using simple machine learning model.

The function of information entropy is to quantify the information change process of the local area. The information entropy of one image is actually the expected value of all information saved in this image. Therefore, the change of information entropy in continuous frames is one special form of information gain. The information gain is often used in decision tree algorithm to select characteristics [25]. Actually, the random forest algorithm proposed based on decision tree is still popular in many fields now [26]. These works prove that the information theory is useful for finding the effective characteristics of different classes. Therefore, the ability of information theory in the classification model is not to be questioned. In this paper, the feature matrix used in the training of model is obtained by calculating the information entropy of sensitive area. These feature matrixes can better express the deep characteristics of image and improve the generalization ability of the model.

3. Method

The method proposed in this paper is to achieve the object detection task in a complex environment. The video processing in complex environment is usually more difficult. Thus a new method in this paper is proposed on the basis of information theory. In the method, the information entropy is utilized to quantify the information change process of the local area. The quantify result can well indicate the different nature between the object and the noise. As shown in Figure 1, the focus of the method is the extraction of sensitive areas in the video frame and then the feature matrix is extracted by information entropy calculation to classify sensitive areas.

3.1. Effective Information Map

The effective information represents the valuable information that is used for identifying foreground objects in the frame of video. There is a greater probability of having targets in the area where effective information is assembled in the image. The effective information map can be used to extract sensitive areas of the image, such that the target can be found faster by less computation. Experiments show that about 1000-2000 candidate areas can be found out from a video frame in the pre-process of region-based convolutional networks [14]. However, extracting sensitive areas through effective information map can greatly reduce the number of these candidate areas.

In selective search model, the candidate area set is composed of local regions that are merged after image segmentation [2]. Therefore, the angular points and edges that are important during the extracting processing of candidate areas often represent effective information. Since the Harris corner detection algorithm [21] can distinguish the flat regions, the edge regions and the corner regions in the image, this algorithm can be used as the extraction method of effective information.

The Harris corner detection algorithm can obtain the response value of the corner information through the transformation of corner response function. The response value of the edge region is negative. The response value of the corner region is a larger positive number and the response value of the relatively flat region is a smaller positive number. Assume that the non-flat region in the video frame is the area containing effective information, so the transformation formula of the effective information map is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2018 doi:10.20944/preprints201801.0160.v1

$$I_{\text{info}} = f(x,y) = \begin{cases} 255, & |dst| > (\alpha \cdot \text{Max}(|dst|)) \\ 0, & |dst| \le (\alpha \cdot \text{Max}(|dst|)) \end{cases}$$

dst represents the corner detection result of the video frame. α times of the maximum value of the corner detection result is the threshold to divide the flat region and the non-flat region, i.e., the ineffective information region and the effective information region. Finally, the effective information map can be obtained by the Gaussian filtering and simple morphological processing. As shown in Figure 2, using the Harris corner detection algorithm can filter out a large number of flat regions, such as lake surface, grass lawn and so on. Then, the sensitive area in video frame can be clearly found out from the effective information map obtained through subsequent processing.

(a) (b) (c)

Figure 2. (**a**) Original Image (**b**) Harris Image (**c**) Effective Information Map

3.2. Sensitive Area Extraction

Using the context information of video is very important to improve video processing efficiency. The frame difference and the background difference are common methods of using context information. In these methods, the consecutive frame of video rather than a single frame is processed and calculated as a basic unit. The dynamic background often contains complex information. So the video with dynamic background is generally difficult to be processed by using the background difference method. The frame difference method is characterized by low complexity, fast running speed and strong adaptive ability of dynamic environment. Some noise in the dynamic background can sometimes be mistaken for the foreground object during the processing of the frame difference method. So this paper tries to improve the frame difference method to avoid the influence of noise as far as possible.

According to the effective information map rather than the original video frame as the processing unit, the frame difference method can get better result. The method used in this paper is to calculate on the basis of effective information maps of three consecutive frames. The calculation formula is

$$D_n(x,y) = [f_n(x,y) - f_{n+1}(x,y) \wedge f_n(x,y)] \vee [f_n(x,y) - f_n(x,y) \wedge f_{n-1}(x,y)]$$

$f_n(x,y)$ represents the effective information map of the n'th video frame. The $\vee$ operation in the formula is to calculate the mean value of corresponding pixels in two effective information maps. The $\wedge$ operation in the formula is to reserve the corresponding pixel value of the effective information map of n'th video frame. This method can retain the outline of the moving object well after removing the background area. The threshold processing is required for $D_n(x,y)$ after the difference operation. The threshold value T_{Otsu} is obtained by the Otsu [24] method automatically. Then the influence of light fluctuation is added on this basis to get the optimal threshold value T_{optimal}. The optimal threshold calculation formula is

$$L_{\text{Diff}} = \frac{1}{N_A} \sum_{(x,y)\in A} (|f_{n+1}(x,y) - f_n(x,y)| + |f_n(x,y) - f_{n-1}(x,y)|)/2$$

$$T_{\text{optimal}} = T_{\text{Otsu}} + \lambda \cdot L_{\text{Diff}}$$

λ represents the influence factor of light fluctuation in the current environment. Through the threshold processing, the difference image can be obtained finally.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2018 doi:10.20944/preprints201801.0160.v1

The sensitive area is the area where the foreground object is easy to appear in video frame. Sensitive areas can be determined easily in the difference image of consecutive video frames. The difference image obtained after threshold processing is divided into many small regions. These regions have the same size and they are non-overlapping. The number of effective information pixel in each region represents the occurrence possibility of the foreground object. This occurrence possibility is the basis for judging sensitive areas. This paper assumes that the region is judged as sensitive area as long as its occurrence possibility is greater than zero. The acquisition of sensitive areas can greatly save the computing cost of subsequent operations.

3.3. Sensitive Area Screening

The extraction of sensitive area can effectively save the calculation cost in the subsequent operation. The another key of the method in this paper is how to select real target areas from all of the sensitive areas. In the most recent period, a large number of end-to-end machine learning models are used to recognize targets [11]. But not all classification problems need to be solved by the deep network [22]. In the NoScope model, just a simple convolutional neural network is used to recognize objects that belong to a small number of categories [4]. The training of deep network often requires a large number of sample labels. During the running process, the deep network also consumes a lot of hardware resources. In the video which contains a dynamic background, these shortcomings of the deep network are easy to be magnified because of many noise of the background.

In the video with complex background, the location recognition of the foreground object is important than the classification of the foreground object. During the consecutive video frames, the information changes of the target area are often progressive. Intuitively, the information changes of the noise area tend to have the characteristics of step change. Therefore, this paper proposes a method that uses information entropy to quantify the process of information change of the area. Then, a simple binary classification model can be trained through these information entropy matrixes. Since the model is based on information change of local area, rather than image area itself, to distinguish the foreground object from the background noise, the model has the characteristics of simple, fast and strong robust.

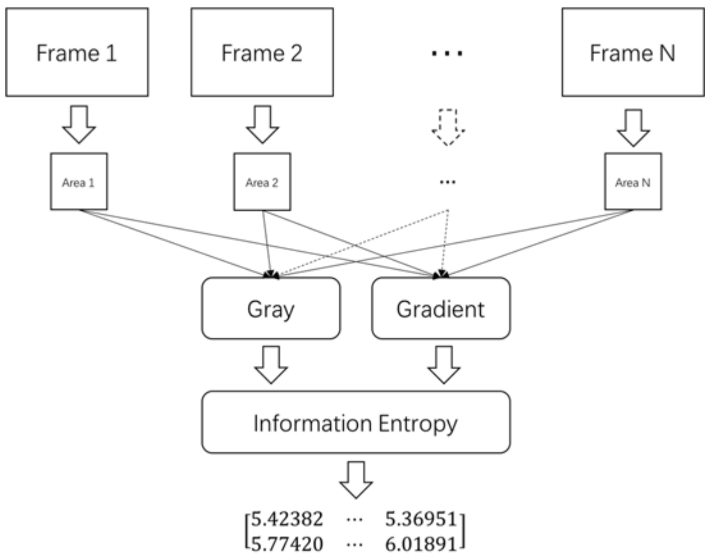
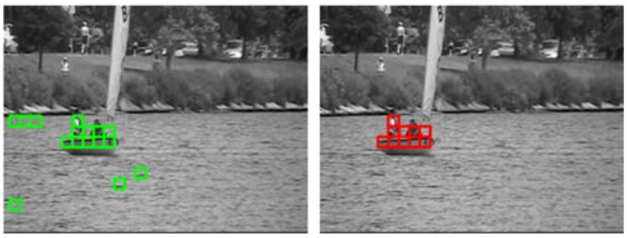


Figure 3. The extraction flow chart of feature matrix of information entropy.

The extraction process of information entropy feature of the sensitive area is shown in the Figure 3. An N dimensional feature vector can be obtained by calculating the information entropy of local areas of the same location in consecutive N frames. But the local areas of video frame are firstly preprocessed by M kinds of algorithms before the calculation of feature vector. Thus, one area in the video frame can be finally represented by a M×N dimensional feature matrix. There are many options for image processing algorithm in preprocessing operation. Two preprocessing algorithms used in this paper are image gray algorithm and image gradient algorithm. No matter which kind of methods, it can be of more generality after the calculation of information entropy.



(**a**) (**b**)

Figure 4. (**a**) Sensitive Areas (**b**) Target Areas

During the training of the machine learning model, one feature matrix is used as the sample data of one sensitive area. These samples are used for training a supervised binary classification model. As shown in Figure 4, this model is used for screening sensitive areas in the video frame, in other words, filtering out the noise generated by dynamic background.

3.4. Target Object Locating

The target object is calibrated on the basis of target areas in the video frame. Since the previous operation has divided the whole image into multiple regions, the result image which contains all target areas is resized for the convenience of subsequent processing. The adjustment rule of image size is that each target area corresponds to one pixel in the adjusted image. Through the morphological analysis of the adjusted image, target areas which belongs to the same object are merged into a large target block. After the image is adjusted back to its previous size, the four boundaries of this irregular target block can be utilize to form a complete rectangle. Finally, these rectangular boxes can be used as the calibration results of the target objects in the video frame. In the next moment, they can be used for classification or other more significant work. The method that uses target areas for locating the target object is not complicated, but it can achieve a nice result.

4. Experiment

The experiment data of this paper is from the video dataset in CDnet 2014 [23]. The main test set used in this paper is the video with dynamic background. The ground truth image in dataset only distinguish the foreground object from the background. Thus, the detection task of this paper mainly focuses on the calibration of the target locations, not on the recognition of the target category. One of the main contributions of this paper is to use a more efficient algorithm to achieve the selection work of candidate areas. The implementation of the regional convolutional network usually leads to excessive running cost because of the time-consuming selection work in it. The end-to-end deep network does not need to screen out candidate areas in advance, so it becomes more popular now. The work in this paper has positive significance to use a low-cost simple model for achieving the same effect with an end-to-end deep network.

The experimental flow chart of the method proposed in this paper is shown in Figure 5. During the transform process of effective information map, the threshold of Harris corner detection algorithm is 0.01. If the shot environment of video is not very complex, extreme or uncommon, then the influence of this parameter is little. During the extraction and screening of sensitive areas, experiments prove that the optimal pixel size of sensitive area is 11×11. After the sensitive area is determined, the same location area of 8 consecutive frames are selected as data sample source for subsequent classification work. Then, each sensitive area can be transformed into a 2×8 dimensional feature matrix.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2018 doi:10.20944/preprints201801.0160.v1

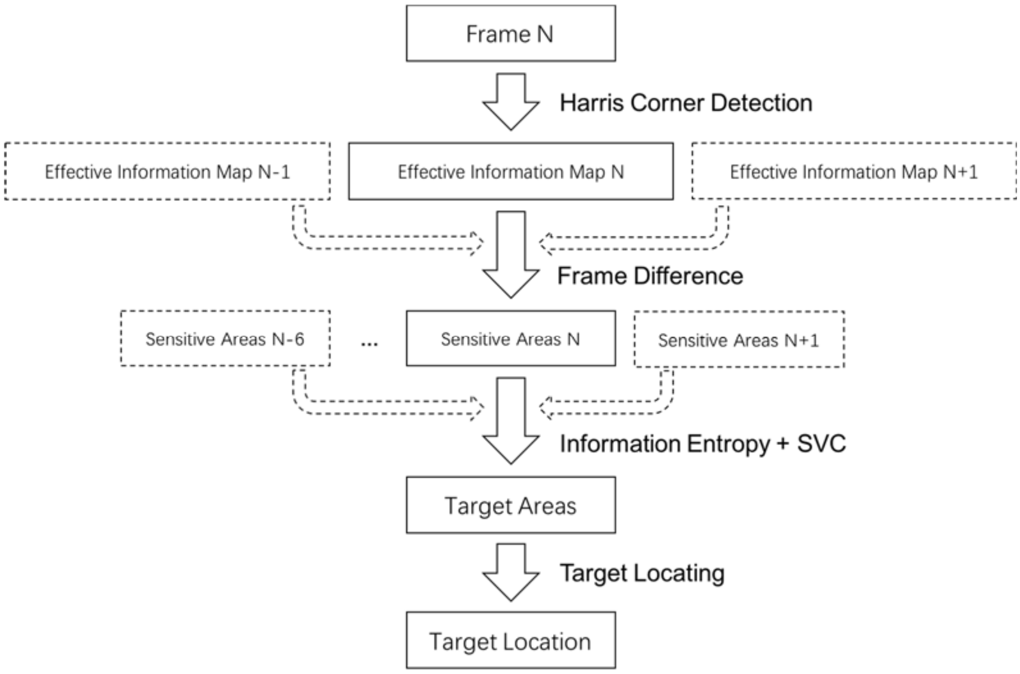


Figure 5. The experimental flow chart of method.

Table 1. The model parameters table.

Kernel	Degree	Gamma	Error Precision
polynomial	3	1/16	0.001

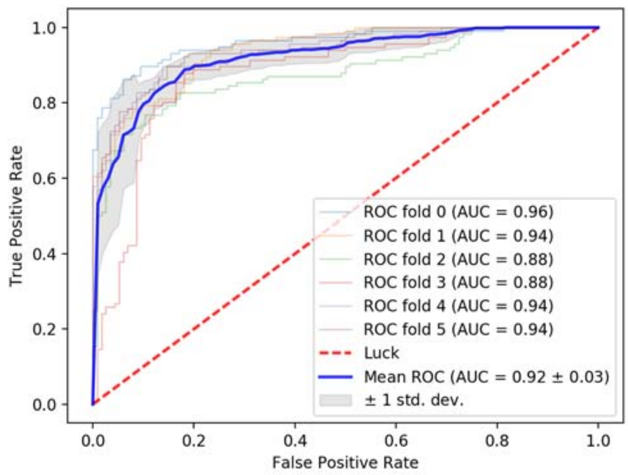
Table 2. The experimental result table.

Video Category	Dynamic Background		Camera Jitter
Video Content	Boats	Highway	Traffic
Accuracy	85.29%	85.61%	88.05%
Speed(fps)	64	52	40

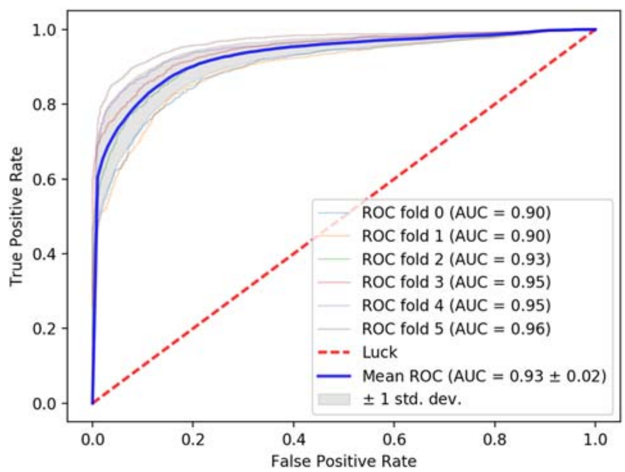
In the next period of the experiment, different classifiers are trained for several video environments respectively. However, the training parameter of different classifiers are the same. The supervised binary classification model used in the experiments is SVC model. The parameters of model are shown in Table 1. Three representative series of video are selected as primary experiment data: one contains the dynamic background of water fluctuation, one contains the dynamic background of leaves shaking and the other one contains heavy jitter of the camera. As shown in Table 2, the experimental results prove that the object detection framework based on information entropy can achieve the same ideal effect in different scenes. After running the different classification models with cross-validation, the ROC curves are plotted as shown in Figure 6. It can be proved that the object detection framework proposed in this paper has the strong adaptive ability in different scenes.

The different classification models based on information theory are trained in different scenes. Then, these models are tested in the corresponding scene. The test results are shown in Figure 7. In the dynamic scene which content is a boat, about 1000 frames in the first half of the video are used in the training of the detection model and about 2000 frames in the second half of the video are used in the test. In Figure 7, the first line is the processing result obtained by using the validation set of video and the second line is the processing result obtained by using the test set of video. If the classification model is trained based on the image itself rather than the information entropy result, the model may get high accuracy in validation set but will be followed by poor generalization ability in the test set.

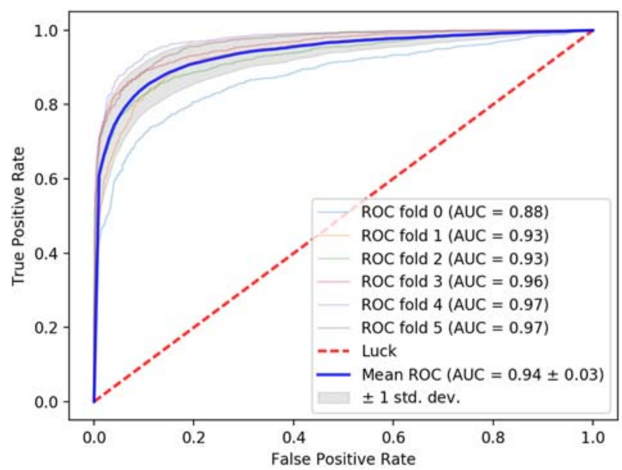
The above experiment proves that the video detection framework based on information theory has strong generalization ability in the same scene.



(**a**)



(**b**)



(**c**)

Figure 6. The ROC curves. (**a**) Boats (**b**) Highway (**c**) Traffic

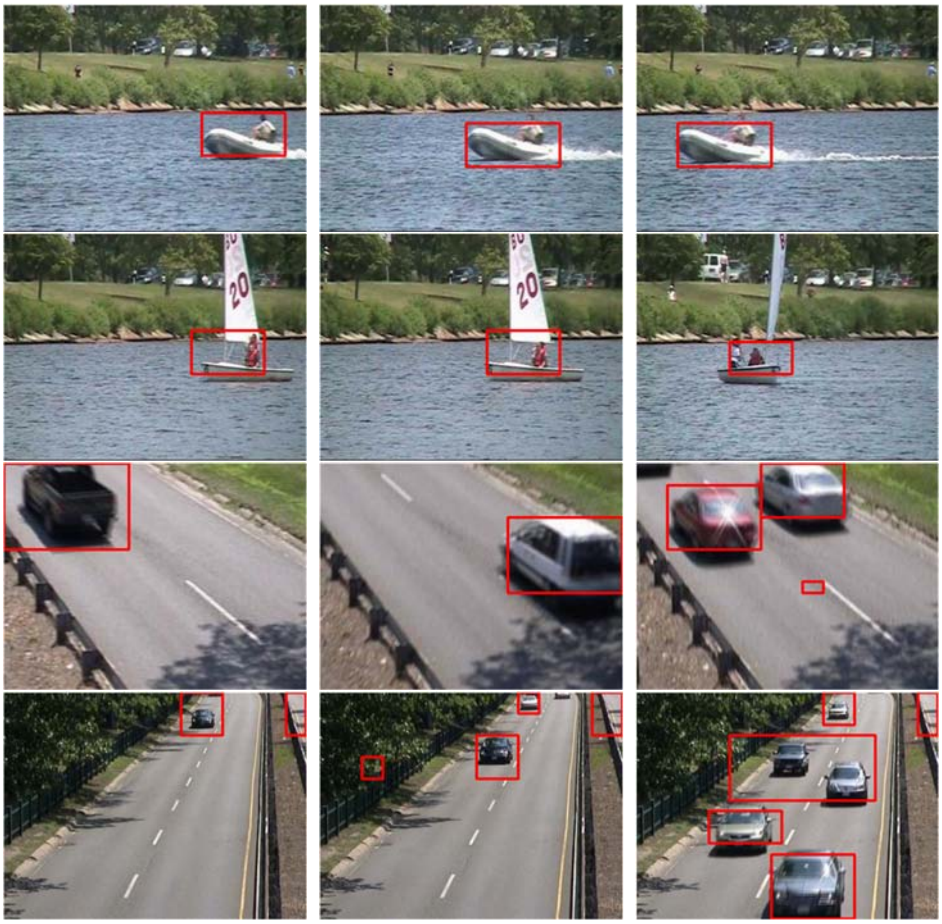


Figure 7. The test results.

5. Conclusion

A new target detection method based on information theory is presented in this paper. Although this method does not have the ability to classify target categories, the method can be treated as a framework for target location detection to save the cost as much as possible. A large number of candidate areas are avoided to be considered by using the context information of the video. At the same time, much hardware cost can be saved by using a lightweight machine learning model. The method of this paper provides a new way, which is different from the end-to-end deep network, to solve the video processing task at low cost.

Although the algorithm has strong generalization ability in the same scene, different models need to be trained respectively for video with different dynamic backgrounds, because the noise has different characteristics in different dynamic background. Thus the future work is to find the common characteristics among the different scenes by using information theory.

6. Acknowledgment

This work is supported by the National Key Research and Development Program of China (No.2016YB0700502) and the National Natural Science Foundation of China (No. 51532006).

References

1. Huang J, Rathod V, Sun C, et al. Speed/accuracy trade-offs for modern convolutional object detectors[J]. 2016.
2. Uijlings J R, Sande K E, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.
3. Zhu X, Dai J, Yuan L, et al. Towards High Performance Video Object Detection[J]. 2017.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2018 doi:10.20944/preprints201801.0160.v1

4. Kang D, Emmons J, Abuzaid F, et al. NoScope: Optimizing Neural Network Queries over Video at Scale[J]. Proceedings of the Vldb Endowment, 2017, 10(11):1586-1597.
5. Zhu X, Xiong Y, Dai J, et al. Deep Feature Flow for Video Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017:4141-4150.
6. Lowe D G. Object Recognition from Local Scale-Invariant Features[C]// iccv. IEEE Computer Society, 1999:1150.
7. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.
8. Zhu Q, Yeh M C, Cheng K T, et al. Fast human detection using a cascade of histograms of oriented gradients[C]// Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006:1491-1498.
9. Verschae R, Ruiz-Del-Solar J, Correa M. A unified learning framework for object detection and classification using nested cascades of boosted classifiers[J]. Machine Vision & Applications, 2008, 19(2):85-103.
10. Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object detection with discriminatively trained part-based models[J]. Computer, 2014, 47(2):6-7.
11. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2014, 115(3):211-252.
12. Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[M]// AB initto calculation of the structures and properties of molecules /. Elsevier, 2015:85-91.
13. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
14. Girshick R, Donahue J, Darrell T, et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(1):142.
15. He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1904-16.
16. Girshick R. Fast R-CNN[J]. Computer Science, 2015.
17. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.
18. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015:779-788.
19. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[J]. 2016.
20. Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015:21-37.
21. Harris C. A combined corner and edge detector[J]. Proc Alvey Vision Conf, 1988, 1988(3):147-151.
22. Ba L J, Caruana R. Do Deep Nets Really Need to be Deep?[J]. Advances in Neural Information Processing Systems, 2013:2654-2662.
23. Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, CDnet 2014: An Expanded Change Detection Benchmark Dataset, in Proc. IEEE Workshop on Change Detection (CDW-2014) at CVPR-2014, pp. 387-394. 2014.
24. Otsu N. A Threshold Selection Method from Gray-Level Histograms[J]. IEEE Transactions on Systems Man & Cybernetics, 2007, 9(1):62-66.
25. Quinlan J R. C4.5: programs for machine learning[J]. 1993, 1.
26. Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.