**Visual Tracking with Deep Neural Network-based**

**Object Detection and Dynamic Image Masking**

Dong-Hyun Lee

Dept. of IT Convergence Engineering, Kumoh National Institute of Technology

61 Daehak-ro, Gumi, Gyeongbuk, Republic of Korea

donglee@kumoh.ac.kr

**Abstract**

**This paper proposes a visual tracking algorithm for videos with deep neural network-based object detection framework and dynamic image masking with heuristic tracking strategy. In order to consider spatio-temporal constraints of object movement in a video, the proposed algorithm estimates the pixel velocity of the bounding boxes in a sequence of previous frames, sets the search region (SR) for the current frame, and masks the rest of the area in the frame for the object detector to focus on a single object that it tracks. To take account of the motion smoothness of the object, the algorithm selects the best candidate from the new bounding box candidates by considering their confidence scores as well as the distance from their centers to the center of the previous bounding box. The proposed algorithm is tested on VOT(Visual Object Tracking) datasets, which is a standard tracking benchmark to demonstrate its performance.**

Keywords-component; visual tracking; deep neural network; dynamic masking

**I. Introduction**

In recent years, convolutional neural network based approach has been widely used in image classification such as AlexNet [1] and ResNets [2], as well as object detection such as Faster R-CNN [3], and R-FCN [4]. However, they are only applicable to still images for classification and object detection and cannot be directly applied to visual tracking in video since the still images and the video images have different properties.

Tracking an object in a video is a challenging problem due to motion blur, occlusion, illumination change, and size change of moving objects in a video. For this reason, object detection algorithms for still images cannot be directly applied to the visual tracking task. There have been numerous works on visual tracking such as GOTURN [5] and MDNet [6]. However, they use rather simpler network than image-based object detection algorithms for fast feature learning. Moreover, most of them only consider the object detection box of the previous frame to estimate the location of the object in the current frame, which limits taking advantage of spatio-temporal constraint of the moving object in a video.

In this paper, a novel visual tracking algorithm based on the the state-of-the art object detector is proposed to robustly track a single moving target with high accuracy and low failure. Instead of developing a network only for tracking, the prposed algorithm enables the general purpose object detector to be used for visual tracking without any modification. This is advangageous since there are numerous image detection benchmark with large amount of dataset such as ImageNet [7] and PASCAL VOC [8]. Although there is a couple of benchmark datasets for visual tracking such as VOT [9] and YouTube Object Dataset (YTO) [10], the size of the datasets are not as large as thouse of image datasets. The algorithm consists of the convolutional neural network based object detector, dynamic masking, and heuristic tracking strategy. For the object detector, YOLOv2 [11] is used, which is one of the fastest algorithm. The dynamic masking enables the general purpose object detector for images to track the moving object in a sequence of video with less disturbance by the other objects. Unlike the other approaches where only crop the area of the previously estmated box from the current image frame, the proposed dynamic masking uses three bounding boxes, i.e., the estimated box before last, the previously estimated box, and the predicted bounding box. For the algorithm to predict the current location of the target, the pixel velocity of previously detected boxes are used. The predicted box is defined by shifting the previously detected box by the pixel velocity. Moreover, instead of cropping, masking is used to only maintain the area where the tracking object is likely to be located in the image frame. This enables the proposed algorithm to take advantage of spatio-temporal constraint of moving object in a video. Although the dynamic masking reduces disturbance of the object detector by the objects in the same class as the tracking object beloings to, it still suffers from losing track of the target object if the other objects are not clearly masked by the dynamic masking. Therefore, the heuristic tracking strategy is used for the object detector to focus on the tracking object. The strategy first considers the confidence probability of the detected objects in the current frame to remove low confidence detection. It then sums up three distances, i.e, the distance between each candidate and the center of the previously estimated box, the distance between the candidate and the center of the predicted box, and the distance between the candiate and the line that passes through the centers of the two consecutive previous boxes. The heuristic strategy then selects the candiate that has the smallest sum of the distances as the final target box in the current frame.

This paper is orgainzed as follows. In Sec II, the proposed visual tracking is introduced. In Sec III, the experimental result is described. Finally, Sec IV concludes the paper.

**II. Proposed Visual Tracking**

*A. Object Detection Network*

In the proposed tracking algorithm, YOLOv2, which is one of the fastest CNN-based object dector, is used as the state-of-the art object detection algorithm. YOLOv2 divides the input image into S by S grid cells and each cell contains five anchor boxes with different sizes (width and height). In the training, the cell that contains the center of the target object is in charge of training its anchor boxes to determine the class of the object as well as to find the x and y offset of the object center position and the size of the object. Each anchor box predicts the confidence probability and the class probabilities as well as the box center and size.

Since the proposed algorithm aims to detect and track a single and specific target in real-time instead of detecting multiple objects with different classes, it is preferable to use small size network with faster performance. For this reason, the proposed algorithm uses Tiny-YOLOv2, which is a simpler version of the original one that has less layers for faster object detection. There are two types of implementation version of YOLOv2, the Darknet version and the the Google Tensorflow library version. In this paper, the Tensorflow version of Tiny-YOLOv2 is used to implement the proposed algorithm [12]. It has total eight layers consisting of convolutional layers and maxpooling layers as shown in Table I. The intial sizes of the five anchor boxes are (1.08, 1.19), (3.42, 4.41), (6.63, 11.38), (9.42, 5.11), and (16.62, 10.52). The configuration of the network and the initial box sizes can be downloaded from [12]. The filter size of the last layer is 30 since there are 5 anchor boxes with 5 coordinates (x, y, width, height, confidence) and 1 class (target class).

TABLE I

FULL DESCRIPTION OF TINY-YOLOv2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Layer | Type | Filters | Size / Stride | Output |
| 1 | Conv. Maxpool | 16 | 3 x 3  2 x 2 / 2 | 416 x 416 |
| 2 | Conv. Maxpool | 32 | 3 x 3  2 x 2 / 2 | 208 x 208 |
| 3 | Conv. Maxpool | 64 | 3 x 3  2 x 2 / 2 | 104 x 104 |
| 4 | Conv. Maxpool | 128 | 3 x 3  2 x 2 / 2 | 52 x 52 |
| 5 | Conv. Maxpool | 256 | 3 x 3  2 x 2 / 2 | 26 x 26 |
| 6 | Conv. Maxpool | 512 | 3 x 3  2 x 2 / 1 | 13 x 13 |
| 7 | Conv. | 1024 | 3 x 3 | 13 x 13 |
| 8 | Conv. | 30 | 3 x 3 | 13 x 13 |

*B. Dynamic Masking*

The dynamic masking aims for the object detection network to robustly keep track of it with fast speed by masking area in the image where the target object is less likey appear. In oreder to do this, the search region is defined as a union of sequential bounding boxes. The first box is the estimated box before last, denoted as , which is two frames ahead from the current frame. The second box is the previously estimated box, . The third box is the predicted box, , which is defined by shifting the center of , denoted as , by adding vector , which is defined as the difference between and :

. (1)

The predicted box enables the detector to consider the area where the target is likely to be located based on the pixel velocity of previously estimated bounding boxes. The three boxes are padded by half of their sizes and the search region at frame *t,* *SR*(*t*) is defined as the union of them:

. (2)

From *SR*(*t*), each pixel of the frame *t* is masked to zero if the pixel location is not in *SR*(*t*). The dynamic masking takes advantage of the spatio-temporal consistency of the target movement and is able to guide the general-purpose object detector to focus on the target object with less distractions.

*C. Heuristic Tracking Strategy*

The dynamic masking guide the detector to avoid detecting the objects of the same class of the target object . However, if the non-target objects are within the search region, it is unavoidable for the detector to select one of non-target object that has higher confidence probability than that of the target object. Thus, it is required to provide some heuristic tracking strategy for the detector to be biased on the tracking target.

The proposed heuristic tracking strategy consists of two phases. The first phase utilizes the confidence probabilities of the target candidates to filter out low confidence objects. There are two thresholds, one is the global threshold, denoted as , which is used to filter out less likely objects. The other one is the local threshold, denoted as , which is defined as the half of the highest confidence probability of each frame. Only the candidates where their confidence probabilities are higher than are considered at the second phase. In the second phase, the strategy determines the target which has the smallest sum of the Euclidean distances between the center of the candidate and , as well as the distance from the line that passes through and . The selected candidate center, denoted as , is defined as

(3)

where

.

The function represents the Euclidean distance between the pixel point *a* and *b*, and the function *line*(*a*, *b*) is the line function that passes through *a* and *b*. From (3), the estimated box, denoted as , whose center is is selected as the final estimated box at frame *t*.

TABLE II

TEST RESULTS ON VOT DATASET

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Pedestrian | | Gymnast | |
| Avg. IOU | # of Failures | Avg. IOU | # of Failures |
| YOLOv2 | 0.273 | 148 / 567 | 0.448 | 13 / 306 |
| Proposed | 0.344 | 13 / 567 | 0.442 | 1 / 306 |

**III. Experiments**

The proposed algorithm was compared with the YOLOv2 by testing the tracking performance of pedestrian and gymnast. For training pedestrian, VOT 2014 (woman) and VOT 2017 (crossing, frisbee, girl) dataset were used, and for testing human tracking, VOT 2014 (jogging), which is shown in Fig. 1(a), was used. In the case of training gymnastics, VOT 2016 (gymnastics2) and VOT 2017 (gymnastics3 and 4) were used, and for testing gymnaistics tracking, VOT 2017 (gymnastics2) shown in Fig. 2(a) was used. For the tracking performance metrics, an average of IOU (Intersection Over Union) and the number of failures were used.Table I shows the test results of the pedestrian and the gymnast of YOLOv2 and the proposed algorithm. The result from the pedestrial tracking shows that the proposed algorhm tracked the target more precisly with less failures. In te case of the gymnast tracking, the average IOUs of YOLOv2 and the proposed algorithm were almost the same. However, the number of failures of the porposed algorithm was much less than that of YOLOv2.

Fig. 1(b) and (c) (Fig. 2(b) and (c))show the test results of YOLOv2 and the proposed algorithm, respectively, with the input frames in Fig. 1(a) (Fig. 2(a)). In the figure, the white, red, cyan boxes are the ground truth, previous, and the estimated boxes, respectively. The green and the yellow boxes are the estimated box before last and the predicted boxes, respectively. As shown in the figure, while YOLOv2 failed tracking when there was motion blur due to the fast movement of the target object or when occlusion happens, the proposed algorithm was able to track the target robustly. The overall results in Table I, Fig. 1 and Fig. 2 demonstrate that the proposed algorithm is able to track a target with less failures than YOLOv2.



Fig. 1 The VOT 2014 (jogging) dataset for testing (a). The tracking results from YOLOv2 (b) and the proposed algorithm (c).

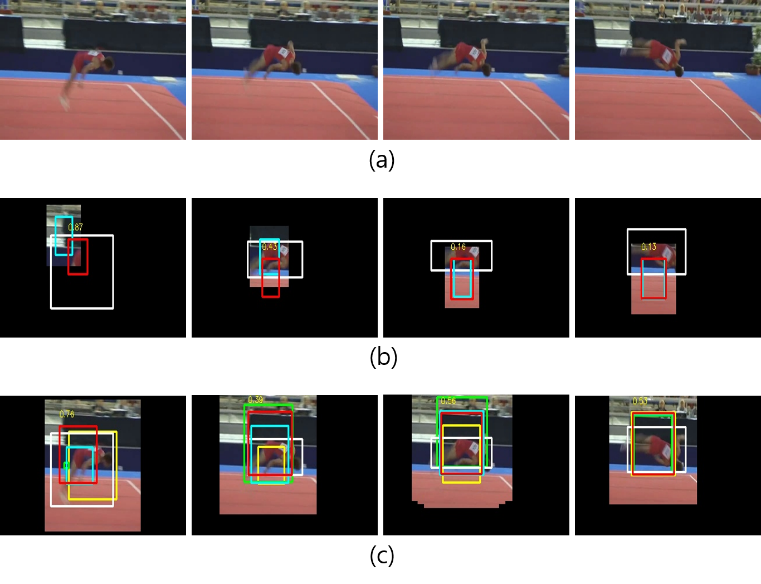


Fig. 2 The VOT 2014 (jogging) dataset for testing (a). The tracking results from YOLOv2 (b) and the proposed algorithm (c).

**IV. Conclusion**

In this paper, the object detection-based visual tracker is proposed. The algorithm efficiently combines the state-of-the art object detector with dynamic masking and the heuristic tracking strategy for visual tracking. In the dynamic masking, the three bounding boxes, i.e., the estimated box before the last, the previous box, and the predicted box, are used for dynamic masking. The heuristic tracking strategy enables the object detection network to robustly track a moving target with motion blur and occlusion.

**Acknowledgment**

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-2014-0-00639) supervised by the IITP(Institute for Information & communications Technology Promotion)

**References**

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91-99.

[4] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 379-387.

[5] D. Held, S. Thrun, and S. Savarese, “Learning to track at 100 fps with deep regression networks,” in *European Conference on Computer Vision*, 2016, pp. 749-765.

[6] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293-4302.

[7] O. Russakovsky et al, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, pp. 211-252, 2015.

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International journal of computer vision, vol. 88, pp. 303–338, 2010.

[9] M. Kristan et al., “The visual object tracking vot2015 challenge results,” in *IEEE international conference on computer vision workshops*, 2015, pp. 1-23.

[10] A. Prest, C. Leistner, J. Civera, C. Schmid, and Y. Ferrari, “Learning object class detectors from weakly annotated video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3282-3289.

[11] J. Redmon, and A. Farhadi, “YOLO9000: better, faster, stronger,” arXiv preprint arXiv:1612.08242, 2016.

[12] Darkflow, Available: https://github.com/thtrieu/darkflow