International Journal of Internet, Broadcasting and Communication Vol.13 No.2 156-165 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.2.156

IJIBC 21-2-20

Object Tracking with Histogram weighted Centroid augmented Siamese Region Proposal Network

Sutanto Edward Budiman† and Sukho Lee††

†*Supersell Co. Ltd, Researcher, Korea*
†† *Professor, Dept. Information Communications Engineering, Dongseo University, Korea*
E-mail petrasuk@gmail.com

Abstract

In this paper, we propose an histogram weighted centroid based Siamese region proposal network for object tracking. The original Siamese region proposal network uses two identical artificial neural networks which take two different images as the inputs and decide whether the same object exist in both input images based on a similarity measure. However, as the Siamese network is pre-trained offline, it experiences many difficulties in the adaptation to various online environments. Therefore, in this paper we propose to incorporate the histogram weighted centroid feature into the Siamese network method to enhance the accuracy of the object tracking. The proposed method uses both the histogram information and the weighted centroid location of the top 10 color regions to decide which of the proposed region should become the next predicted object region.

Keywords: *Siamese network, Histogram, Deep learning, Object Tracking.*

1. Introduction

Most advanced countries have strengthened their public security level by applying object tracking systems to surveillance cameras. Object tracking systems can be used by security applications to detect or track desired objects such as people. However, there are still many challenges in video object tracking such as illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, and low resolution[1].

The main objective of video object tracking is to locate the target region in the next frame which corresponds to the object target in the previous frame. This is done by comparing the patch regions in the previous and the next frames and locate the most similar patch in the next frame. Nowadays, many good deep learning based object detection and re-identification methods have been proposed which can re-target an object in a new frame [1,2,5,6,7,8,9,10]. The difference between video object tracking and object detection is that in object tracking, the object that we are trying to locate can be any target in the patch, but in object detection, the object is that

Manuscript Received: March. 17, 2021 / Revised: March. 20, 2021 / Accepted: March. 22, 2021
Corresponding Author: petra@g.dongseo.ac.kr
Tel: +82-51-320-1744, Fax: +82-51-327-8955
Professor, Dept. Information Communications Engineering, Dongseo University, Korea

which the network has learned from given labels. If we apply an object detection network to the object tracking problem, the tracking will show problems when there is appears a non-target object that looks similar to the target object we want to track. To avoid this problem, we should give a good method which can be used to calculate the similarity between the same object in the previous frame and the current frame.

2. Related Works

Siamese networks is consist of two identical network structure that are used to compare an original patch from the previous frame with a search region. Original patch is define as cropped target object in the previous frame. However the search regions refers to a larger cropped image which is centered based on the previous frame. In object tracking, Siamese network is used because of its processing speed and can maintain good accuracy.

Region proposal network (RPN) was introduced in Faster R-CNN[2] in 2015. Region proposal network consists of two network. One of the network called box-classification layer (cls) which is used to differentiate between the background and foreground classification, and other network is called box-regression layer (reg) which is used to make a proposal for targeted object[]. In region proposal network used anchors which are 3 different size boxes with 3 ratios (*i.e.* 1:1, 1:2, 2:1). In other words region proposal network will use 9 different anchor.

SiamRPN[3] was introduced by Bo,Li *et al*, which combine the Siamese network with the region proposal network (RPN). Feature extraction done by the Siamese subnetwork and used in the RPN. RPN in SiamRPN made from two networks, one is used to distinguish between the background and the foreground, while other one make the proposal regions for the target object.

In Online Information Augmented SiamRPN[4], the SiamRPN is used to generate meaningful candidates and make measurement using histogram and distance between two center point for each set of candidates in the current frame with the next frame. However, this method is based on simple color based online information where the positional relationship between the colors are not fully regarded. Therefore, in this paper, we propose the use of the histogram weighted centroid as the online information to be used with the SiamRPN to get better tracking results.

3. Proposed Method

SiamRPN has good performance between distinguishing the foreground and background. In the Figure 1 shows the activation maps is large values for all person shown in the patch but not in the background, this happen because they are belong in the same person class. Figure 1 shows that SiamRPN is capable for creating good candidates in this problem, and also it is possible that the activation will switch from the targeted object to close non-targeted object that will cause failure result in the tracking.

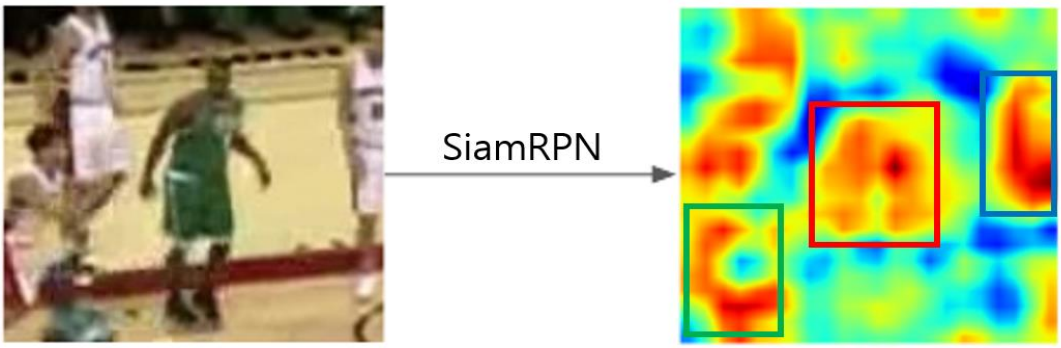

Figure 1. Showing the top-3 regions including large activation values generated by the SiamRPN

Instead of discarding the large activation regions that SiamRPN produce, we propose a method that can process this activation maps to prevent failure in tracking.

We denote the regions proposed by the SiamRPN by $S_1, S_2, .., S_T, \ldots S_N$. These regions are generated by the SiamRPN method. We can use all the proposed regions, or some selected regions, like, the top-3 regions which have the largest activation value in the proposal regions. Among all the proposal regions, we determine the tracking object region to be that which has the largest summation value of the normalized frequency of pixels divided by the difference of the centroids between the current and the next frames:

$$S_T{}^* = \max_{S_T} \sum_n \frac{{F_n}^k}{||{c_n}^k - {c_{T,n}}^{k+1}||^2 + 0.001} \tag{1}$$

where $S_T{}^*$ is the final tracking region, F_n^k denotes the *n*-th color bin of the tracking target in the current frame, and c_n^k denotes the histogram weighted centroid of the tracking target in the current frame, which is calculated by the following equation:

$${c_n}^k = \left(\frac{\sum_n {F_n}^k {x_n}^k}{\sum_n {x_n}^k + 0.001} \;,\; \frac{\sum_n {F_n}^k {y_n}^k}{\sum_n {y_n}^k + 0.001}\right). \tag{2}$$

Here, ${x_n}^k$ and ${y_n}^k$ are the average values of the x and y coordinates corresponding to the *n*-th color bin of the *k*-th target region, and the summation with respect to *n* is for the top-10 frequent color bins. Meanwhile, ${c_{T,n}}^{k+1}$ denotes the histogram weighted centroid of the T-th target region in the next$(k+1)$ frame. There are as many ${c_{T,n}}^{k+1}$ values as the number of target regions and there are calculated as follows:

$${c_{T,n}}^k = \left(\frac{\sum_n {F_n}^k {x_{T,n}}^k}{\sum_n {x_{T,n}}^k + 0.001} \;,\; \frac{\sum_n {F_n}^k {y_{T,n}}^k}{\sum_n {y_{T,n}}^k + 0.001}\right). \tag{3}$$

Here, the calculation of F_n^k is the same for both c_n^k and ${c_{T,n}}^{k+1}$as we want to track the same colors that appeared in the target region in the current frame. Therefore, the top-10 frequent color bins used in Eq. (1) are all those in the target region in the current frame.

The histogram weighted centroid of the each color bin in the target object includes the information on the

location of the colors in the target region, and also includes information which are the major colors to be tracked since it is weighted by the histogram. Therefore, in the case that the SiamRPN cannot well discriminate between the tracking target and the false objects, it can further help to discriminate between the target and false objects. The SiamRPN and the histogram weighted centroid information mutually compensate for each others since the SiamRPN will focus more on the overall shape of the target object, while the histogram weighted centroid will further help to discriminate the colors and the positional relationships between the colors, a kind of information that deep learning based tracking methods can not discriminate well since they are trained on many tracking objects with various colors.

Figure 2 shows the flowchart of the proposed tracking method. Both the proposal region extraction and the histogram weighted centroid information extraction work very fast, so that the combination of both techniques works in real-time with various sizes of video sequence frames.

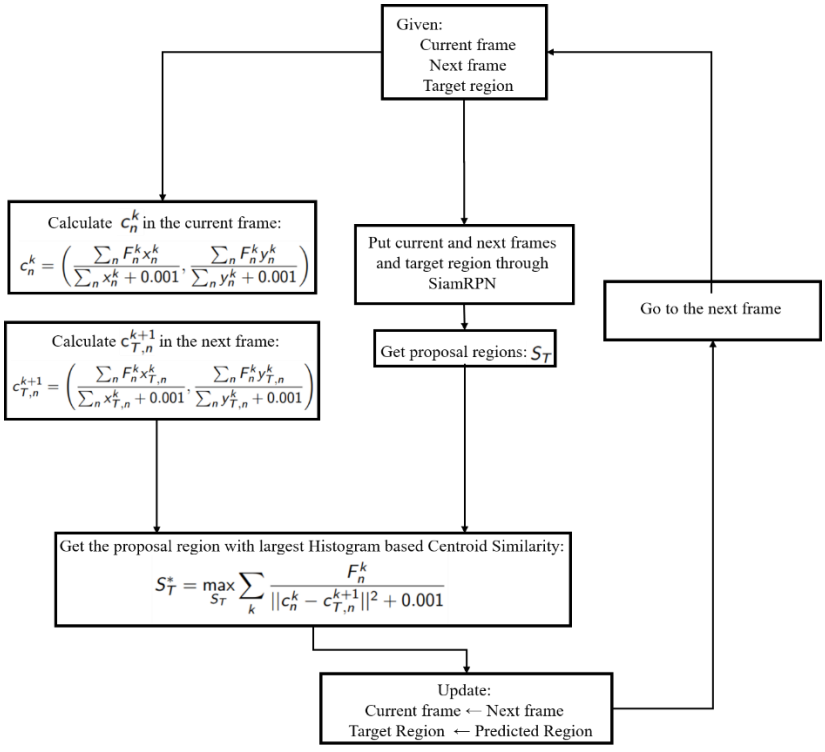

Figure 2. Flowchart of the proposed method

4. Experimental Results

We compared the proposed method with several other methods, i.e., with the SCM[5] (Sparsity-based Collaborative Model), the STRUCK[6] (Structured Output Tracking with Kernels), the TLD[7] (Tracking-Learning-Detection), the ASLA[8] (Adaptive Structural Local Sparse Appearance Model), the CXT[9] (Context Tracker) model, the VTS[10] (Visual Tracker Sampler) model. Table 1 describes the abbreviation of the proposed method and the SiamRPN, where we used two different versions of the histogram bins: One with a 4-bin histogram (OURS4HISTO) and another with an 8-bin histogram (OURS8HISTO) used in the calculation of the centroids. The ONSI method is a method where a simple online information has been used together with the SiamRPN (not the histogram weighted centroid). We included the method in [4] to show the validness of the histogram weighted centroid information. We use OTB benchmark dataset [1] as our test

dataset for our tracking method. OTB benchmark are most used for benchmarking tracking results. Trackers are compared with area under curve (AUC) of success rate for one pass evaluation (OPE). We use the IoU(Intersection over Union) measure to evaluate the tracking performance.

Table 1. Method Name

Method	Name
ONSI[4]	Online Information Augmented SiamRPN
SiamRPN	Original SiamRPN
OURS4HISTO	Proposed with 4 histogram bins
OURS8HISTO	Proposed with 8 histogram bins
Other Methods	Provided by Object Tracking Benchmark

We experimented the tracking performance on various situations and show the results in the success plot figures below. Figure 3(a) shows the overall tracking result on the whole OTB benchmark dataset. Figure 3(b) shows the success plot under different illumination variation environment, where the illumination in the target region is significantly changed for different frames. Figure 3(c) shows the success plot for situation where the variation of the ratio of the bounding boxes in the first frame and the current frame is large(ratio of 2).

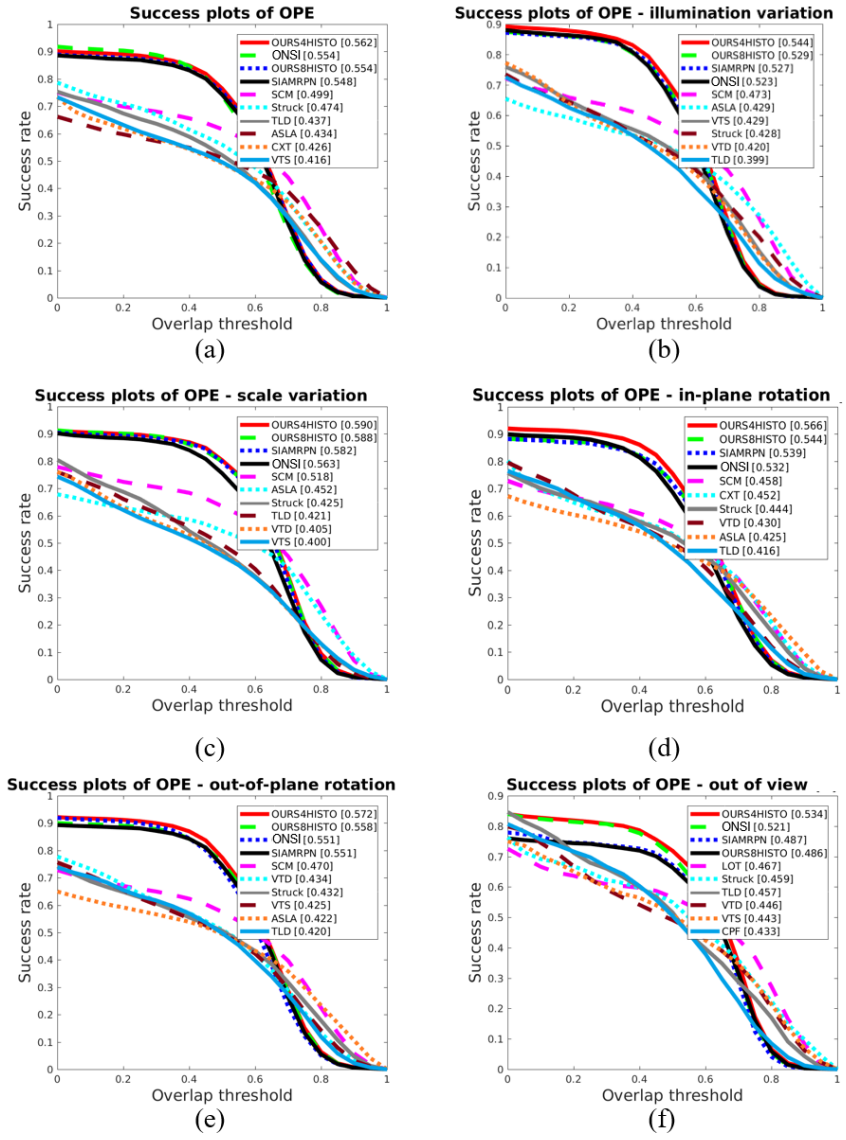

Figure 3. Tracking performance under (a) normal (b) illumination variation (c) scale variation (d) in-plane rotation (e) out-of-plane (f) out-of-view situation

Figure 3(d) to Fig. 3(f) show the situations where the target has in-plane-rotation (the target rotates in the image plane), out-plane-rotation(the target rotates out of the image plane), and out-of-view(some portion of the target leaves the view) situation, respectively. Figure 4(a) to Fig. 4(c) show the success plots when the target is partially or fully occluded, has a non-rigid object deformation, and is blurred due to the motion of target or camera, respectively. Finally, Fig. 4(d) to Fig. 4(f) show the success plots when the image contains fast motion (motion larger than 20 pixels), background clutters, and low resolution. As can be seen from the figures, the proposed method shows the best performance in the tracking success with IoU $\leqq$ 0.6. A level of IoU=0.6 is actually large enough for the tracking object to be said to be detected so we can say that the proposed method shows a good tracking performance.

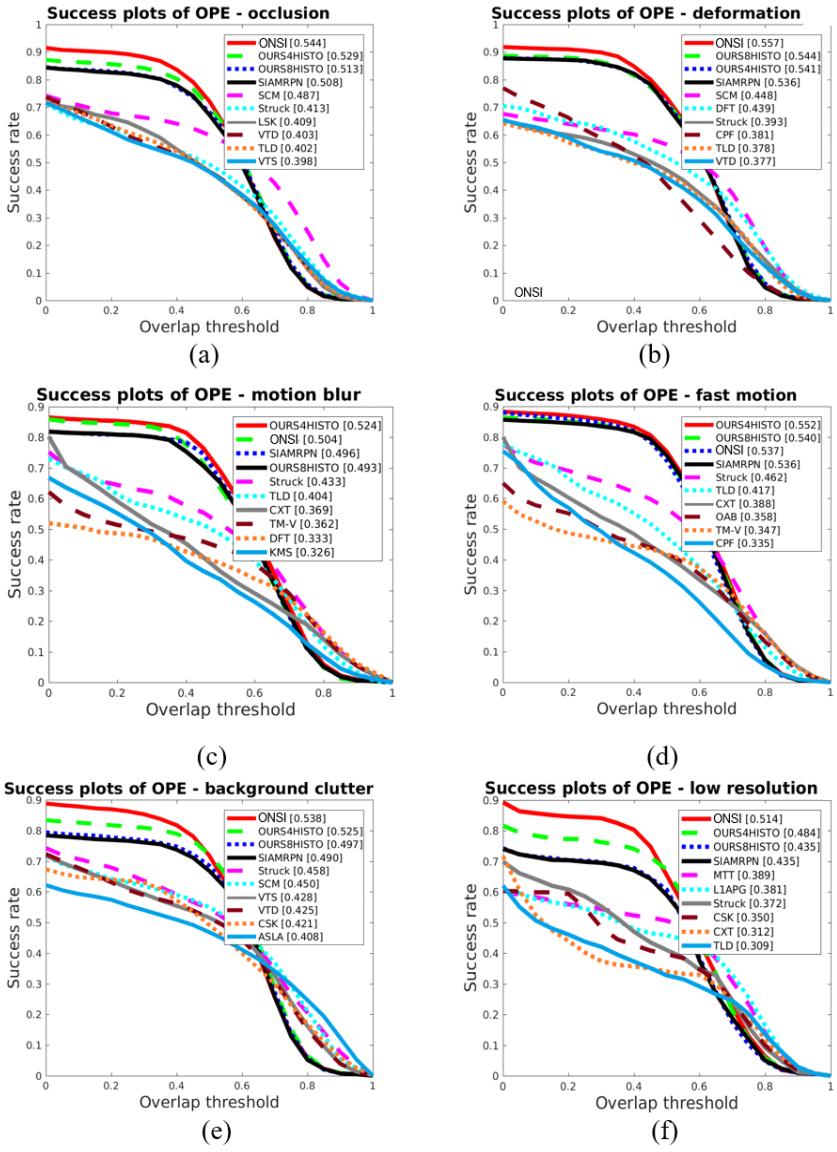

Figure 4. Tracking performance under (a) occlusion (b) deformation (c) motion blur (d) fast motion (e) background clutter (f) low resolution situation

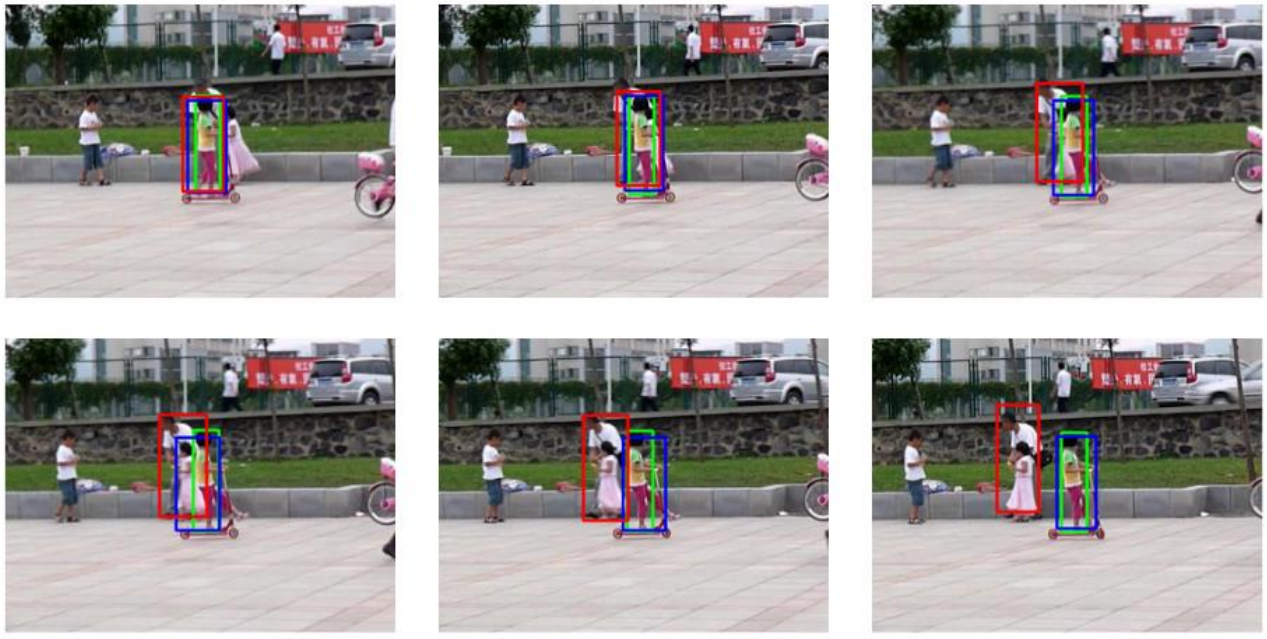

Figure 5. Comparison between the original SiamRPN method (red box) with proposed method (blue box) on the pedestrian video sequence. Green box is the ground-truth box.

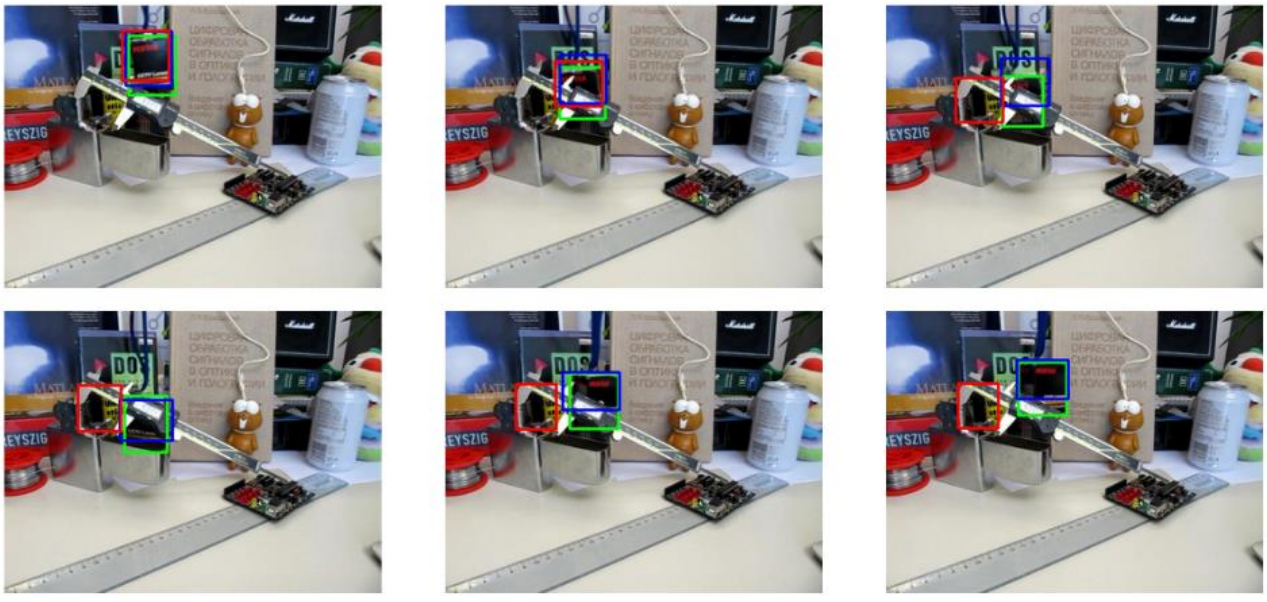

Figure 6. Comparison between the original SiamRPN method (red box) with proposed method (blue box) on the in-door video sequence. Green box is the ground-truth box.

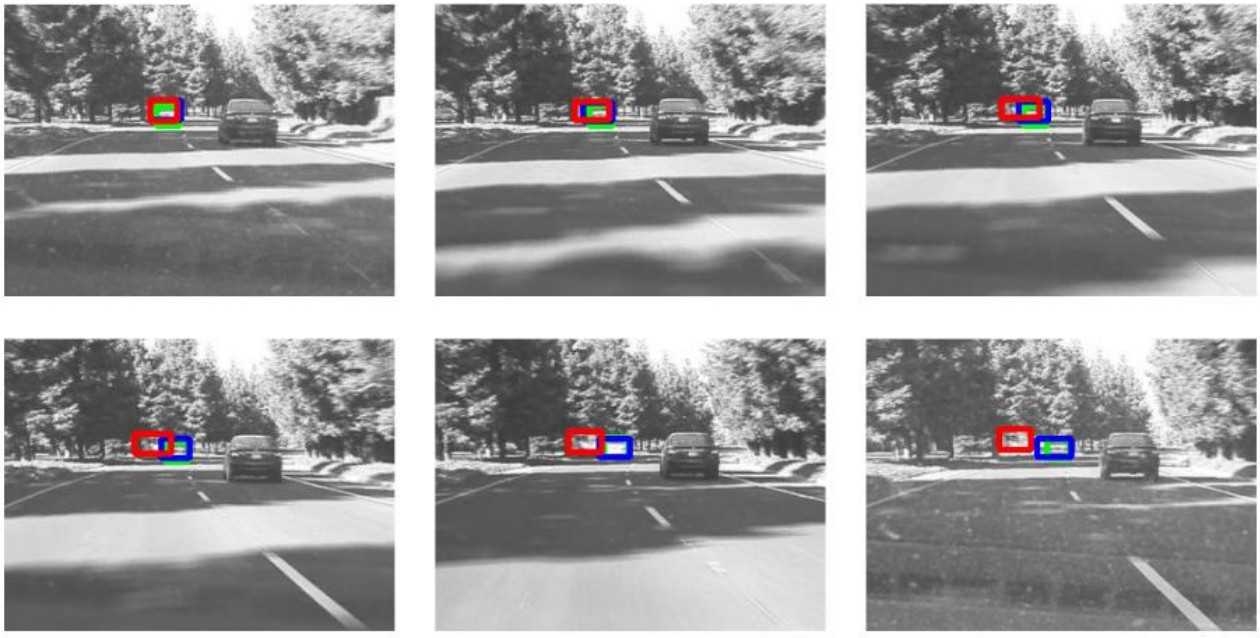

Figure 7. Comparison between the original SiamRPN method (red box) with proposed method (blue box) on the vehicle video sequence. Green box is the ground-truth box.

Figure 5-7 compare the tracking results with the original SiamRPN to validate the use of the histograms weighted centroid information. It can be seen that the original SiamRPN fails to track the object in several situations when the background has similar colors to the target object or when partial occlusion appears. However, with the proposed method, the target object is tracked robustly as can be seen in the figures.

5. Conclusion

In this paper we propose method that enhances the performance of tracking based on Siamese network which is represented by SiamRPN. Using SiamRPN to generate several candidates and apply the histogram weighted centroid method to finally choose between the proposed regions to determine the final target region. The combined information of the histogram weighted centroid and the proposed region of the SiamRPN work together well so that the target object can be tracked well even under several difficult situations. Further studies which combine the online information with deep learning based networks should be explored to enhance the performance of tracking methods.

ACKNOWLEDGEMENT

This work was supported by the Technology development Program(S2840023)funded by the Ministry of SMEs and Startups(MSS, Korea).

References

[1] Y. Wu, J. Lim, H. Yang, “Object tracking benchmark,” IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9), 1834–1848 (Sep 2015). https://doi.org/10.1109/TPAMI.2014.2388226

[2] S. Ren ,K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards realtime object detection with region proposal networks," in Neural Information Processing Systems 28, pp.91–99, Dec.7-24, 2015.

[3] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, "High performance visual tracking with siamese region proposal network," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8971–8980, Jun.18-23, 2018. DOI: https://doi.org/10.1109/CVPR.2018.00935

[4] E.B. Sutanto, S. Lee, "Online Information Augmented SiamRPN," in 12th International Conference on Computer Vision Systems, pp.480-489, Sep.23-25, 2019.

[5] W. Zhong, H. Lu, M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in IEEE Conference on Computer Vision and Pattern Recognition, pp.1838-1845, Jun.16-21, 2012. DOI: https://doi.org/10.1109/CVPR.2012.6247882

[6] S. Hare, A. Saffari, P.H.S. Torr, "Struck: Structured output tracking with kernels," in International Conference on Computer Vision, pp.263-270, Nov.6-13, 2011. DOI: https://doi.org/10.1109/ICCV.2011.6126251

[7] Z. Kalal, J. Matas, K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.49-56, Jun.13-18, 2010. DOI: https://doi.org/10.1109/CVPR.2010.5540231

[8] X. Jia, H. Lu, M.-H Yang, "Visual tracking via adaptive structural local sparse appearance model," in IEEE Conference on Computer Vision and Pattern Recognition, pp.1822-1829, Jun.16-21, 2012. DOI: https://doi.org/10.1109/CVPR.2012.6247880

[9] T.B. Dinh, N. Vo, G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in CVPR, pp.1177-1184, Jun.20-25, 2011. DOI: https://doi.org/10.1109/CVPR.2011.5995733

[10] J. Kwon, K.M. Lee, "Tracking by Sampling Trackers," in International Conference on Computer Vision, pp.1195-1202, Nov.6-13, 2011. DOI: https://doi.org/10.1109/ICCV.2011.6126369