SCALING AND OCCLUSION ROBUST ATHLETE TRACKING IN SPORTS VIDEOS

Jianghu Lu[1], *Di Huang*[2,*], *Yunhong Wang*[1], *and Longteng Kong*[1]

[1]State Key Lab. of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
[2]IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

ABSTRACT

This paper proposes a novel approach to athlete tracking in sports videos. It follows the framework of Compressive Tracking (CT), but extends it by two manners, *i.e.* scale refinement as well as occlusion recovery. For the former, an objectness method, namely Edge Box (EB) is adopted to generate proposals, replacing the fixed sampling box in CT, which better fits the scales of the candidate objects. For the latter, a candidate obstruction based solution is presented, which makes use of additional trackers to detect possible obstructions especially the ones possessing highly similar appearances as the target one, and relocate the target as occlusion ends. Therefore, the proposed method inherits the advantage of CT in robust object modelling and fast processing speed, and embodies the tolerance to occlusion and scaling. We evaluate the proposed method on a collection of videos of beach volleyball games, and the experimental results and the comparison with recent advanced trackers highlight its effectiveness.

Index Terms— Sports Video Analysis, Object tracking, Occlusion, Scaling

1. INTRODUCTION

Sports video analysis has received increasing attention both in academia and industry in recent years for its scientific challenges and promising applications. It involves in a large variety of research directions, containing data statistics, highlight extraction, content insertion, computer assisted referee, tactics analysis, *etc*. Among these directions, athlete tracking is a major issue, which plays a fundamental role for automatic processing. Object tracking aims to locate a moving object (or multiple objects) over time in the video, and it is a hot topic within the domain of computer vision and intelligent surveillance. The past decade has witnessed its progress in many applications, such as traffic monitoring, pedestrian detection and Human-Computer Interaction (HCI). Due to the complexity and diversity in the real-world condition, various approaches to object tracking are proposed in literature, and they generally include two principal modules, *i.e.* appearance representation and model update. The former one is how to comprehensively describe the target object, including holistic template [1, 2], sparse representation [3, 4] and discriminative model [5, 6]. The latter one lies in accurately capturing the appearance change as the object moves, *e.g.* template update [7], online boosting [8] and incremental subspace update [9]. A number of approaches claim that they are competent at handling the cases in the public datasets, *e.g.* VIVID [10], CAVIAR [11], but the video clips are only composed by limited (hundreds of) frames, which makes them problematic in practice.

Different from the benchmark videos for general object tracking, the sports video has some specific difficulties, which make it even more challenging to track the athlete. On the one hand, the athlete is frequently occluded, to which the majority of the existing approaches are quite sensitive. Indeed, several solutions have been investigated to improve the robustness to occlusion. For instance, [12, 13] detect occlusions by using the depth clue of the tracked object; [14, 15, 16] deal with it through a combination of the motion and appearance of the object by linear or non-linear dynamic models; and [17, 18] handle this problem with the help of multiple cameras. However, the occlusion cases are not as tough as those in sports videos, where the players (object and obstruction) often share high similarity in appearance, including height, figure, dressing, *etc*, leading to confusion in the current solutions. Furthermore, these methods usually require more clues, *e.g.* in the depth modality or from additional cameras, which are not always available in the given situation. On the other hand, the players in the sports video move rapidly everywhere within the filed of view, incurring large scale changes. A straightforward solution is to vary the size of the candidate box so that it best fits the scale of the object, but a fine tuning step needs more computational expenditure while a coarse one tends to fail when the scale drastically changes, where the balance is hard to achieve. An alternative is to use scale invariant features [19], such as SIFT, nevertheless, it also greatly increases the time cost and the template size has to be sufficiently large to extract enough features for matching. Such facts suggest that the solutions are not applicable in this task.

A few attempts have been made to track the athlete in sports videos. [20] tracks the players in basketball and hockey games from the view of tactics analysis, and the authors try to predict all the possible moving directions for players, but it may incur failure for infinite possibilities. [21, 22] use parti-

* indicates the corresponding author (dhuang@buaa.edu.cn)

cle filters to predict the position and velocity of the players in beach volleyball games. They separate foreground and background to make athlete modeling easier, but they lose many cues in background to improve stability and accuracy.

In this paper, we propose a novel and powerful approach to track the athlete in a given sports video captured by a single static camera. It is based on Compressive Tracking (CT) [23], but significantly improves it in two aspects, *i.e.* robustness to occlusion as well as insensitivity to scaling. To deal with occlusions, a candidate obstruction that moves towards the target is considered to relocate the object as they separate from each other, which in particular works when the object and obstruction are similar. To handle the problem caused by scaling, we replace the box sampled by CT with the proposals detected by Edge Box [24], each of which best fits the size of the possible object in it, for similarity measurement between the object and candidate in multi-scale image feature space. Therefore, the proposed method inherits the advantage of CT in robust object modelling and fast processing speed, and embodies additional tolerance to occlusion and scaling variations. We evaluate the method on a collection of videos of beach volleyball games, the experimental results and the comparison with recent advanced trackers highlight its effectiveness.

2. SCALING AND OCCLUSION ROBUST CT

Compared with general object tracking, athlete tracking in sports videos suffers from severe scale changes and frequent occlusion variations, and processing speed is also an important indicator. Considering its decent performance and computational simplicity, we follow the framework of Compressive Tracking (CT) for athlete tracking in sports videos, and extend it to deal with the aforementioned problems. For completeness, we review CT and describe the solutions to scaling and occlusion subsequently.

2.1. Compressive Tracking

CT, recently proposed in [23], is an effective and efficient method, which formulates the tracking problem as a detection task. It uses an appearance model based on features extracted in the compressed domain, and thus combines the advantages of generative and discriminative approaches.

In appearance representation, a set of multi-scale discriminative features are selected using information-preserving and non-adaptive dimensionality reduction based on the theory of compressive sensing. A small number of randomly generated linear measurements prove sufficient to retain most of salient information and achieve good reconstruction of the signal if it is compressible, thus allowing efficient projection of the original feature space to a low-dimensional compressed subspace.

For model update, when the object is initialized at the first frame, positive candidates near the current location and negative ones far away from it are sampled at the next frame, and the candidate with the maximal similarity score is used to determine the current location and update the classifier.

Specifically, to capture appearance changes in real time, positive and negative candidates, denoted as b_{pos} and b_{neg} respectively, are sampled around the current location of the tracked object. Each candidate is represented by a low dimensional feature vector $\mathbf{v} = (v_1, \ldots, v_n)^T$, and all the elements in $\mathbf{v}$ are supposed to be independently distributed and modeled with a naive Bayes classifier as in [25];

$$H(\mathbf{v}) = log\left(\frac{\prod_{i=1}^{n} p(v_i|y=1)}{\prod_{i=1}^{n} p(v_i|y=0)}\right) = \sum_{i=1}^{n} log\left(\frac{p(v_i|y=1)}{p(v_i|y=0)}\right) \quad (1)$$

where we assume uniform prior, $p(y=1) = p(y=0)$. $y \in \{0,1\}$ is a binary variable representing the candidate label, and H denotes the bayesian classifier. (1) is used to score every sampled box, and the one with the highest score is selected as the new location of the target, which updates the classifier simultaneously, denoted as

$$update(H(\mathbf{v}), b_{pos}, b_{neg}) \quad (2)$$

2.2. Refinement for Scaling

In [23], to deal with the scale problem, CT convolves the candidate with a set of rectangle filters at multiple scales. Each filtered image is reshaped as a column vector, and all the vectors are concatenated into a very high-dimensional feature for appearance representation. Such processing provides CT with some robustness to scale changes. However in sports videos, the players often move at a quite high speed, which causes the swift and frequent change in scale, and since the candidate box in CT is of a fixed size, the multi-scale feature based representation cannot always be accurate. What is more, the candidate box is generated by a sliding window within a certain distance, and an elaborate trade-off has to be made between the accuracy and time cost by choosing a proper step length. To overcome these drawbacks, we adopt object proposal techniques which output a number of windows that are likely to contain individual objects, to refine those candidates. In this study, we consider Edge Box (EB) [24] for its high performance and low time complexity. EB first detects the edges within a given region whose size is moderately bigger than that of the object, and then locates all the proposals by counting the number of edge lines in them. In contrast to the candidates cropped by a fixed box in CT, the ones within the produced proposals by EB have different sizes, which better fit the possible objects in them. In the following, similar to CT, the proposals within a certain distance are selected and normalized in size to compare with the current object in the multi-scale feature space, and the one with the maximal likelihood is finally predicted as the next location. Fig.1 demonstrates the process of scale refinement by EB.

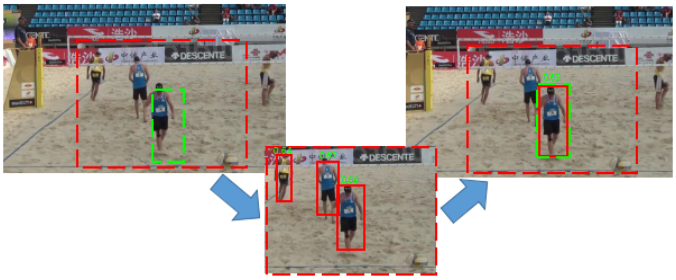

Fig. 1. Process of scale refinement (CT candidate in the green box and EB candidates in the red ones).

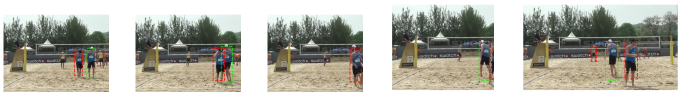

Fig. 2. Process of occlusion recovery (object in the green box and obstruction in the red one).

2.3. Recovery in Occlusion

Due to the use of Haar-like features, CT presents better tolerance to occlusion than some advanced counterparts, *e.g.* MILTrack and Struck. However, as described in Sec.1, occlusion in sports videos is more difficult, because the athletes of high similarity often occlude each other. To handle this issue, we present a novel solution. Considering that before occlusion, there should exist one or more candidate objects approaching the target one, we individually focus on the candidate(s). To be more specific, when the new location of the target is determined, within a relatively larger area, we search the regions with the variance values bigger than a pre-defined threshold Var_{thr}, as they tend to contain moving objects. Based on the scores in classifier H, we further filter out the ones similar to the target object according to a threshold Ret_{thr}, indicating that there are another or more similar objects around the target. For simplicity, in this study, we only concentrate on the candidate with the maximal likelihood:

$$b : \max_{j=1}^{n} H(v_j)$$
$$s.t. \quad \begin{cases} variance(b_j) \geq Var_{thr} & (j = 1, 2, 3, \cdots, n), \\ H(v_j) \geq Ret_{thr} & (j = 1, 2, 3, \cdots, n). \end{cases} \tag{3}$$

There are thus two classifiers to track the target object and candidate obstruction simultaneously. The jaacard distance is adopted to measure the overlap between them, and decide if they occlude each other using a threshold Ove_{thr}:

$$O(b_i, b_j) = \left| \frac{b_i \bigcap b_j}{b_i \bigcup b_j} \right| \geq Ove_{thr} \tag{4}$$

If the candidate obstruction box is above that of the target (the obstruction is nearer to the camera), the object is occluded. In this case, only the candidate obstruction tracker works while the target object tracker stops. Similar to finding the candidate obstruction, the algorithm relocates the target when they separate from each other. Fig.2 illustrates such a process. The whole procedure of this method is given in Alg.1.

Algorithm 1 Scaling and Occlusion Robust CT

Input: frame sequence: $f_1, f_2, \ldots, f_n$, classifier: H, initial object: b_{ini}, frequency: M
Output: object sequence: $b_1, b_2, \ldots, b_n$

classifier: $H, candidate = false, occlusion = false$
for $i = 0$ to n **do**
  if $!occlusion$ **then**
    if $i \,\%\, M == 0$ **then**
      $EdgeBox(f_i, b_{i-1})$
    else
      $SampleBox(f_i, b_{i-1})$
    end if
    $CompressFeature, b_i := max(H(\{v_{ij}\}))$
    if $!candidate$ **then**
      if $variance(box) \geq Var_{thr}, H(\{v_{ij}\}) \geq Ret_{thr}$ **then**
        $C_i := max(H(\{v_{ij}\})), candidate = true$
      end if
    else
      if $overlap(b_i, C_i) \geq Ove_{thr}$ **then**
        $occlusion = true$
      end if
    end if
  else
    $SampleBox(f_i, C_i), CompressFeature$
    if $variance(c_{ij}) \geq Var_{thr}$, $H(\{v_{ij}\}) \geq Ret_{thr}$ **then**
      $b_i := max(H(\{v_{ij}\}))$,
      $occlusion = false, candidate = false$
    end if
  end if
end for

3. EXPERIMENTAL RESULTS

To evaluate the method, we conduct experiments on the Beach Volleyball (BeaVoll) dataset. The database, protocol, parameter tuning, and results are described in the following.

In contrast to general object tracking, there are very limited public benchmarks of sports videos. In this study, we use the BeaVoll dataset, from General Administration of Sport in China. It contains 30 video clips of beach volleyball games, from 80 to 120 seconds. Each video is captured from a game by a camera equipped at the end line of competition terrain, and there hence exist variations in background, illumination, body shape and clothing. The location of the player is manu-

ally labeled at each frame as groundtruth. In our experiment, the videos are downsampled from 1440×1080 to 432×240 in resolution. 15 video clips are used in validation for parameter tuning, and the others for test. Fig.3 shows some samples.

Fig. 3. Some samples in the BeaVoll dataset.

Trackers are usually evaluated using the precision rate or the success rate as the indicator. Precision rate is the average Euclidean distance between the center locations of targets and groundtruths; however, if the tracker loses the object, the output is random and the evaluation is less meaningful. Success rate is the overlap of bounding boxes, which counts the number of successful frames where overlap is larger than the given threshold, and this indicator is always reasonable. Therefore, we adopt the latter as most recent studies do.

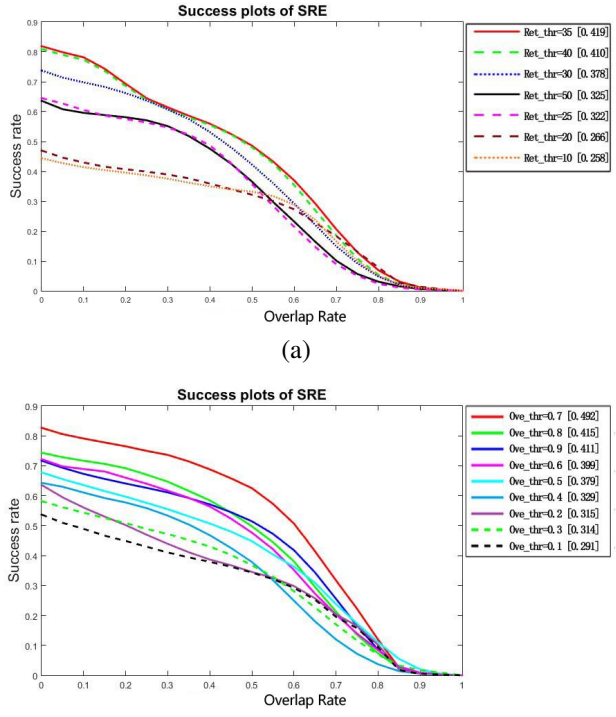

Fig. 4. Parameter tuning based on (a) object retrieval threshold and (b) occlusion overlap threshold.

There are some important parameters crucial to our results, including object retrieval threshold Ret_{thr} and occlusion overlap threshold Ove_{thr}. We set the values according to the accuracy in validation. Fig.4 displays the performance using the two parameters respectively. As shown in Fig.4 (a), when Ret_{thr} is small, even though there is no candidate obstruction, the algorithm still probably mistakes a random box for it. It rejects further detection of true candidate obstructions, thus leading to failure in the subsequent. Similar to Ret_{thr}, Ove_{thr} should be adjusted as well.

We compare the proposed method with the state of the art ones, including CT [23], ASLA [26], CSK [27], DFT [28], ORIA [29], and IVT [9] when varying the overlap rate.

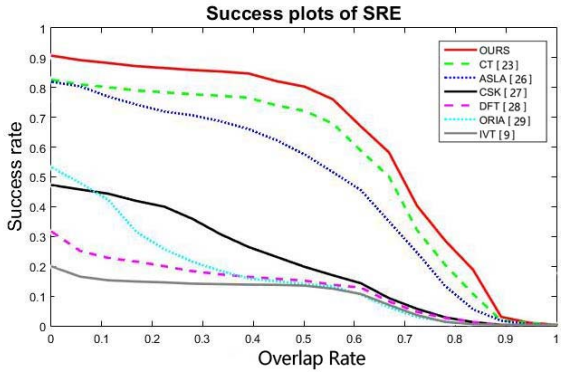

Fig. 5. Comparison between our method and the state of the art ones in terms of success rate on the BeaVoll database.

4. CONCLUSION

In this paper, we propose an effective and efficient method to track athletes in sports videos. It is based on CT, but improves it in two aspects. Scale refinement is achieved by EB based proposal generation and occlusion recovery is reached by introducing the candidate obstruction based strategy. The proposed method is evaluated on the BeaVoll database, and the comparison with the state of the art trackers clearly demonstrates its advantage for this task.

5. ACKNOWLEDGMENT

This work was supported in part by the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China (Grant No. L2015TGA9004), the National Natural Science Foundation of China (Grant No. 61421003, 61573045, and 61540048), and the Beijing Natural Science Foundation (Grant No. 4142032), and the Fundamental Research Funds for the Central Universities.

6. REFERENCES

[1] N. Alt, S. Hinterstoisser, and N. Navab, "Rapid selection of reliable templates for visual tracking," in *CVPR, 2010*.

[2] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *TPAMI, IEEE*, vol. 20, no. 10, pp. 1025–1039, 1998.

[3] M. Xue, H. Ling, W. Yi, E. Blasch, and B. Li, "Minimum error bounded efficient ? 1 tracker with occlusion detection," in *CVPR, 2011*.

[4] Y. Wu, H. Ling, J. Yu, F. Li, M. Xue, and E. Cheng, "Blurred target tracking by blur-driven tracker," in *ICCV, 2011*.

[5] S. Avidan, "Ensemble tracking," *TPAMI, IEEE*, vol. 29, no. 2, pp. 261–271, 2007.

[6] R. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *TPAMI, IEEE*, vol. 27, no. 10, pp. 1631–1643, 2005.

[7] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, 1981.

[8] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting.," in *BMVC*, 2006.

[9] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *IJCV*, vol. 77, no. 1-3, pp. 125–141, 2008.

[10] R. Collins, X. Zhou, and S. K. Tech, "An open source tracking testbed and evaluation web site," in *PETS*, 2005.

[11] R. B Fisher, "The pets04 surveillance ground-truth data sets," in *PETS*, 2004.

[12] M. Harville, G. Gordon, and J. Woodfill, "Adaptive video background modeling using color and depth," in *ICIP, 2001*.

[13] Y. Ma and C. Qian, "Depth assisted occlusion handling in video object tracking," in *ISVC*. 2010.

[14] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *ICCV, 2001*.

[15] Z. Duan, Z. Cai, and J. Yu, "Occlusion detection and recovery in video object tracking based on adaptive particle filters," in *CCDC, 2009*.

[16] D. Tang and Y. Zhang, "Combining mean-shift and particle filter for object tracking," in *ICIG, 2011*.

[17] J. Batista, "Tracking pedestrians under occlusion using multiple cameras," in *ICIAR*. 2004.

[18] M. Mozerov, A. Amato, X. Roca, and J. Gonzàlez, "Solving the multi object occlusion problem in a multiple camera tracking system," *TPAMI, IEEE*, vol. 19, no. 1, pp. 165–171, 2009.

[19] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using sift features and mean shift," *CVIU*, vol. 113, no. 3, pp. 345–352, 2009.

[20] J. Liu, P. Carr, R. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *CVPR, 2013*.

[21] T. Mauthner, C. Koch, M. Tilp, and H. Bischof, "Visual tracking of athletes in beach volleyball using a single camera," *IJCSS*, vol. 6, no. 2, pp. 21–34, 2007.

[22] G. Gomez, P. López, D. Link, and B. Eskofier, "Tracking of ball and players in beach volleyball videos," P-Los, 2014.

[23] K. Zhang, L. Zhang, and M. Yang, "Real-time compressive tracking," in *ECCV, 2012*.

[24] C. Zitnick and P. Dollár, "Edge boxes: locating object proposals from edges," in *ECCV, 2014*.

[25] A Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *NIPS*, vol. 14, pp. 841, 2002.

[26] J. Xu, H. Lu, and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *CVPR, 2012*.

[27] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *ECCV, 2012*.

[28] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *CVPR, 2012*.

[29] Y. Wu, B. Shen, and H. Ling, "Online robust image alignment via iterative convex optimization," in *CVPR, 2012*.