A Survey on Visual Object Tracking: Datasets, Methods and Metrics

V. Ramalakshmi @ Kanthimathi[1], Dr. M. Germanus Alex[2]

[1] *Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India*
[2] *Research Guide, Department of Computer Science, Bharathiar University, Coimbatore, India*

---***---

Abstract: *Object detection and tracking is an important and challenging task in many critical computer vision applications such as automated video surveillance, traffic monitoring, autonomous robot navigation, and smart environments. Object tracking can be defined as the process of segmenting an object of interest from a video scene and keeping track of its motion, orientation and occlusion to extract useful information. Several object tracking methods have been proposed in the past two decades aiming to design a robust object tracker addressing all the practical challenges in the real work. The goal of this paper is to review the datasets, methods, and metrics available in the literature for developing a robust visual object tracker. Particularly, we present the details about five publicly available datasets, twenty object tracking methods, and three metrics for comparing the performance of the visual object tracking systems.*

Key Words: video Processing; object identification; object tracking, visual object tracking, predictive models, and performance metrics.

1. INTRODUCTION

Visual object tracking is one of the challenging problems in the field of computer vision. It has several related applications such as automated video surveillance, traffic monitoring, security and robotics [1]. The objective of a visual object tracker is to estimate the location of a target in all the frames of a video sequence based on the given initial location (or a bounding rectangle) of the target. Object tracking problem has been studied by the computer vision community for several decades. But, still it remains a challenging task to design an efficient and robust visual object tracking system for all the practical real world applications. Further, there are several factors that affect the performance of the object tracker such as illumination variations, scale variations, occlusions, deformations, motion blur, rotations, and low resolutions [2].

Several object tracking methods have been proposed in the past to address some of aforementioned challenges [3] [4] [5] [6] [7]. They are classified into two broader categories: 1) generative models [4] [5] and 2) discriminative models [6] [7]. Generative models formulate the object tracking task as a searching problem. It learns an appearance model of the target object and then searches the best matching appearance window (or the bounding box) in the video frames. They are also referred as template or subspace matching models. Generally, generative methods model the appearance of the target object and they do not consider the background information even through it is useful for differentiating the target from back-ground. On the other hand, discriminative methods formulate the object tracking task as a classification problem. It attempts to differentiate the target object from background information of each video frame. Unlike generative models, discriminate models use both the target object and background information to learn models that separate the target object from the background. Since the usage of background information is beneficial for object tracking, in general, it is shown that discriminative models outperform generative models [2].

In this paper, we review the existing works on visual object tracking in three different perspectives: datasets, methods, and metrics. Firstly, we present the details about the availability of public datasets and what kinds of challenging attributes are available for each dataset. Secondly, we compare and contrast the various visual object trackers proposed in the literature. Finally, we summarize what kinds of performance metrics have been used to evaluate the existing visual object tracking systems. Analyzing all these three perspectives is very crucial for developing an efficient and robust object tracking system.

While there are other churn prediction surveys available in the literature, they primarily focused on different modelling techniques. To the best of our knowledge, none of those surveys reviewed the datasets and metrics for evaluating the churn prediction models. Hence, we believe that this survey can provide a roadmap for researchers to better understand the domain and challenges in detail.

The rest of the paper is organized as follows. In Section 2, we present the components of a model visual object tracking system in detail. In Section 3, we present the details and characteristics about the five publically available datasets for experimenting the visual object tracking system algorithms. In Section 4, we present the details about existing algorithms proposed in the literature for visual object tracking, and compare their performance. In Section 6, we present the details about three performance metrics which are used in the literature for comparing the performance of the visual object tracking algorithms. Finally, in Section 6, we conclude the paper.

2. VISUAL OBJECT TRACKING SYSTEM

In this section, we present the components of a model visual object tracking system in detail.

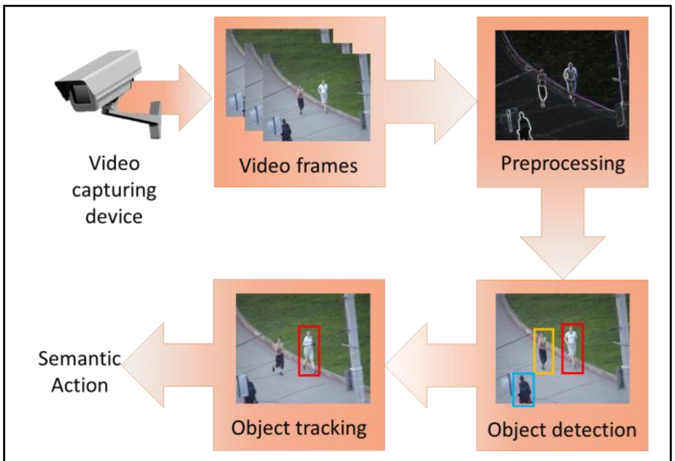

Fig 1. Components of a visual object tracking system

Figure 1 illustrates the different components of a visual object tracking system. It consists of four steps: 1) Data (video frames) acquisition, 2) Preprocessing, 3) Object detection, and 4) Object tracking. In the first step, the object tracking system acquire the video frame from the underlying video capturing devices such as Closed-Circuit Television (CCTV) or web camera. In the second, they system applies some preprocessing methods, such as noise filters, for enhancing the captured video frames. In the third step, the system detects the objects of interest from the given first video frames. This step involves applying back-ground subtraction techniques. The basic idea of background subtraction is subtracting a predefined background model frame from the current from to identify the objects of interesting. The system may identify multiple interested objects. Several advanced background subtraction methods have been proposed in the literature which is insensitive to external environmental conditions such as noise. These include approximate median, running Gaussian, and mixture Gaussian methods.

After detecting the objects of interest, the final step is called object tracking which involves finding the location of the target object in the subsequent video frames. While tracking the object, the system admin shall take semantic actions, based on the type and location of the objects. For example, in case of a traffic monitoring system, the admin can identify and monitor the list of vehicles which are violating the traffic rules and impose fine automatically.

3. DATASETS

In this section, we present the details about the list of existing public datasets and summarize their attributes.

3.1 ALOV300++

Amsterdam Library of Ordinary Videos for tracking, ALOV++[1], aims to cover as diverse circumstances as possible: illuminations, transparency, specularity, confusion with similar objects, clutter, occlusion, zoom, severe shape changes, different motion patterns, low contrast, and so on. It has 11 standard video sequences frequently used in recent tracking papers, on the aspects of light, albedo, transparency, motion smoothness, confusion, occlusion and shaking camera.

The dataset consists of 315 video sequences. The main source of the data is real-life videos from YouTube with 64 different types of targets ranging from human face, a person, a ball, an octopus, microscopic cells, a plastic bag to a can. The collection is categorized for thirteen aspects of difficulty with many hard to very hard videos, like a dancer, a rock singer in a concert, complete transparent glass, octopus, flock of birds, soldier in camouflage, completely occluded object and videos with extreme zooming introducing abrupt motion of targets.

To maximize the diversity, most of the sequences are short. The average length of these aspects is 9.2 seconds with a maximum of 35 seconds. The total number of frames in ALOV300 is 89364. The data in ALOV300 are annotated by a rectangular bounding box along the main axes of flexible size every fifth frame. In rare cases, when motion is rapid, the annotation is more frequent. The ground truth has been acquired for the intermediate frames by linear interpolation. The ground truth bounding box in the first frame is specified to the trackers. It is the only source of target-specific information available to the trackers.

3.2 BoBoT - Bonn Benchmark on Tracking

The BoBoT[2] dataset comprises of several short video sequences showing arbitrary target objects and annotation files containing information about the objects' positions and sizes. It consists of 12 video sequences and a total of 8201 frames. All video sequences are 320x240 pixels at 25 fps encoded with mpeg2, but there might be other individual formats in the future. A video sequence must comply with only two conditions: first, the original frame rate must be 25 fps or more, second, the bitrate should be high enough to avoid distinct compression artifacts. The target's position and size is stored relative to the frame size with a precision of up to six digits. This upright rectangular designation is

[1] http://www.alov300.org/

[2] https://adaptivetracking.github.io/

Dataset	No. of videos	Ground truth available	Total No. of Frames	Source	Properties / Challenges / Categories /
ALOV300++	314	Yes	89364	Amsterdam Library of Ordinary Videos	Light (33), Surface Cover (15), Specularity (18), Transparency (20), Shape (24), Motion Smoothness (22), Motion Coherence (12), Clutter (15), Confusion (37), Low Contrast (23), Occlusion (34), Moving Camera (22), Zooming Camera (29), Long Duration (10)
BoBoT	12	Yes	8201	Bonn Benchmark on Tracking	moving cam, moving target, rotation, fast direction changes, background changes, scale changes, non-rigid target, outdoor, partial occlusion, similar distractors, full occlusion, viewpoint changes, illumination changes
i-Lids	7	Yes	35000	IEEE AVSS 2007	abandoned baggage (Task 1) and parked vehicle (Task
VOT2015	60	Yes	NA	VOT2015 benchmark	illumination change, object size change, object motion, clutter, camera motion, blur, aspect-ratio change, object color change, deformation, scene complexity, and absolute motion http://www.votchallenge.ne
Visual Tracker Benchmark	100	Yes	NA	Visual tracker benchmark	illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, low resolution

Table-2: Summary of publicly available datasets for experimenting visual object tracking system

defined as the tightest-fitting rectangle covering the whole target object. If the target is not visible in a frame, all values should be "0.0". Frame counting starts with 0.

3.3 i-Lids

This is a dataset for event detection in CCTV footage and is a sub-set of the i-Lids dataset[3]. The events of interest appearing in the dataset are abandoned baggage and parked vehicle.

Below are the detail of this dataset
- Location of recording: various locations in the UK
- Number of sequences: 7
- Total number of images: 35000
- Format of images: 8-bit color MOV
- Image size: 720 x 576 pixels
- Video sampling rate: 25 Hz

[3] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html

Name	Citation	Year	Major Contribution
MOSSE and Regularized ASEF	[9]	2010	Pioneering work of introducing correlation filters for visual tracking
CSK	[14]	2012	Introduced Ridge Regression problem with circulant matrix to apply kernel methods
STC	[15]	2014	Introduced spatio-temporal context information
KCF	[16]	2014	Formulated the work of CSK and introduced multi-channel HOG feature.
CN	[17]	2014	Introduced color attributes as effective features
DSST	[11]	2014	Relieved the scaling issue using feature pyramid and 3-dimensional correlation filter
SAMF	[18]	2014	Integrated both color feature and HOG feature; Applied a scaling pool to handle scale variations
RPAC	[24]	2015	Introduced part-based tracking strategy
RPT	[25]	2015	Introduced reliable local patches to facilitate tracking
LCT	[23]	2015	Introduced online random fern classifier as re-detection component for long-term tracking
MUSTer	[22]	2015	Proposed a biology-inspired framework where short-term processing and long-term processing are cooperated with each other

Table-2: List of major contributing papers using correlation filters for visual object tracking [23]

3.4 VOT 2015

The VOT 2015[4] dataset comprises 60 short sequences showing various objects in challenging backgrounds. The sequences were annotated by the VOT committee using rotated bounding boxes in order to provide highly accurate ground truth values for comparing results. The annotations are stored in a text file with the format:

frameN: X1, Y1, X2, Y2, X3, Y3, X4, Y4

where Xi and Yi are the coordinates of corner i of the bounding box in frame N, the Nth row in the text file.

The bounding box was be placed on target such that at most ~30% of pixels within the bounding box corresponded to the background pixels, while containing most of the target. For example, in annotating a person with extended arms, the bounding box was placed such that the arms were not included. Note that in some sequences parts of objects rather than entire objects have been annotated. A rotated bounding box was used to address non-axis alignment of the target. The annotation guidelines have been applied at the judgement of the annotators.

3.5 Visual Tracking Benchmark

This dataset[5] contains 100 sequences from recent literatures. The sequence names are in CamelCase without any blanks or underscores (_). When there exist multiple targets each target is identified as dot+id_number (e.g. Jogging.1 and Jogging.2). Each row in the ground-truth files represents the bounding box of the target in that frame,

(x, y, box-width, box-height).

In most sequences the first row corresponds to the first frame and the last row to the last frame, except the following sequences:

David(300:770), Football1(1:74), Freeman3(1:460), Freeman4(1:283).

Table 1 summarizes the list of publicly available datasets and their characteristics for evaluating the performance of visual object tracking systems.

4. VISUAL OBJECT TRACKING METHODS

Visual object tracking has been studied extensively with lots of applications. There are a lot of researches being carried out in the area of visual object tracking. In this section, we survey some of important researches carried out in this area in the recent years. In this section, we introduce the approaches closely related to our work. In this section, we review the existing literature which is broadly classified into two broader categories: 1) Generative models and 2) Discriminative models.

J. Kwon et al. [4] proposed a visual tracking framework that searches for an appropriate tracker in each frame among a predefined set of trackers. The tracker space consists of various trackers using appearance models, motion models, state

[4] http://www.votchallenge.net/vot2015/dataset.html

[5] http://cvlab.hanyang.ac.kr/tracker_benchmark/index.html

representation types, and observation types. The most appropriate tracker is selected using Markov Chain Monte Carlo method. Several particle filter based trackers are proposed in the literature [12] [13]. Zhang et al. [12] proposed a multi-task tracking system using a particle filter framework. They model particles as linear combinations of a dictionary of templates that are updated dynamically. Belagiannis et al. [13] proposed a two particle sampling strategy based on segmentation to mitigate the scale and appearance changes of the target object. Segmentation is applied to all the propagated particles in the first step, whereas in the second step only the strongest particle is segmented.

On the discriminative model side, several kernel-based methods are proposed. For example, Hare et al. [3] presented an adaptive visual object tracking framework based on structured output prediction. They used a kernel-based structured support vector machine for performing online learning and adaptive tracking. Wang et al. [6] proposed a metric learning framework for object tracking. Both visual tracking and appearance modelling are performed online in their framework. Further, their formulations can handle multiple objects and occlusions. In contrast with all these methods, we used a discriminative correlation filter which is robust in detecting objects under several challenging scenarios such as occlusions, scale variation, among others.

Table 2 summarizes the list of major contributing papers in the recent literature using correlation filters for visual object tracking.

5. PERFORMANCE METRICS

There are several standard performance metrics proposed in the literature to compare the effectiveness of the different classifiers for churn prediction. These metrics are suitable for analyzing the performance of any model which is built using both balanced and unbalanced dataset. The metrics are described below.

(a) **Center Location Error (CLE)**: It is one of the widely used evaluation metrics for object tracking applications. It is defined as the average *Euclidean distance* between the center locations of the manually labeled ground truths and the tracked targets bounded by a rectangle. The CLE is computed for all the frames in a video sequence separately. Then the CLE of all the frames are averaged to calculate the final CLE for a particular video sequence.

(b) **Distance Precision (DP) and Precision Plot**: The CLE value of some frames can be random when the tracking algorithm fails to track the object of interest. In such scenarios, CLE may not measure the tracking performance correctly. Recently a new metric called Distance Precision has been proposed to measure the overall tracking performance. It is defined as the relative number (or percent) of frames in the sequence where the CLR is smaller than a certain predefined threshold. The threshold value can be chosen based on specific application requirements and the size of the target object. The average distance precision is plotted over a range of thresholds in the *Precision Plot*.

(c) **Overlap Precision (OP) and Success Plot**: It is defined as the relative number of frames where the bounding box overlap (number of pixels) between the ground truth and tracked target is larger than a certain threshold. The overlap score S is defined as below

$$S = \frac{|R_t \cap R_g|}{|R_t \cup R_g|}$$

Where, R_t and R_g are tracked and ground truth bounding boxes, respectively. This overlap score S is calculated for each frame separately. Then the overlap precision of a sequence of frames is calculated as the relative number of successful frames whose overlap score is larger than a certain threshold. The ratio of successful frames over a range of thresholds are plotted in the *Success Plot*.

6. CONCLUSION

In this paper, initially, we introduced the churn prediction problem and the significance of using predictive modeling methods to overcome the problem of customer churn in telecom industry. We surveyed the existing visual object tracking methods in detail and summarized them. Unlike other surveys, which primarily focused only on the techniques and the accuracy of visual object tracking, in this survey we presented the characteristics and challenging features of the existing publicly available datasets and various visual object tracking methods. Finally, we surveyed the list of the commonly used metrics proposed in the literature for evaluating the performance of various visual object tracking systems.

ACKNOWLEDGEMENT

Authors would like to thank all the research communities for making the datasets available for public use.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah. Object Tracking: A Survey. ACM Computing Surveys, 38(4):1–45, 2006.

[2] Wu, Yi, Jongwoo Lim, and Ming-Hsuan Yang. "Online object tracking: A benchmark." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2411-2418. 2013.

[3] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured Output Tracking with Kernels. In ICCV, 2011.

[4] J. Kwon and K. M. Lee, "Tracking by sampling trackers," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1195–1202.

[5] V. Belagiannis, F. Schubert, N. Navab, and S. Ilic, “Segmentation based particle filtering for real-time 2d object tracking,” in Computer Vision–ECCV 2012. Springer, 2012, pp. 842–855.

[6] X. Wang, G. Hua, and T. X. Han, “Discriminative tracking by metric learning,” in Computer Vision–ECCV 2010. Springer, 2010, pp. 200–214.

[7] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, “Object tracking with joint optimiza-tion of representation and classification,” Circuits and Systems for Video Technology, IEEE Transactions on, vol. 25, no. 4, pp. 638–650, 2015.

[8] B. V. K. V. Kumar, A. Mahalanobis, and R. D. Juday, Correlation Pattern Recog-nition, Cambridge University Press, 2005.

[9] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 2544–2550.

[10] Zhang, Lei, Yanjie Wang, Honghai Sun, Zhijun Yao, and Shuwen He. "Robust Visual Correlation Tracking." Mathematical Problems in Engineering 2015 (2015).

[11] M. Danelljan, G. H¨ager, F. S. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” in Proceedings of the British Machine Vision Conference BMVC, 2014.

[12] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, “Robust visual tracking via multi-task sparse learning,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2042–2049.

[13] V. Belagiannis, F. Schubert, N. Navab, and S. Ilic, “Segmentation based particle filtering for real-time 2d object tracking,” in Computer Vision ECCV 2012. Springer, 2012, pp. 842–855.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting thecirculant structure of tracking-by-detection with kernels,” in Computer Vision–ECCV 2012. Springer, 2012, pp. 702–715.

[15] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, “Fast visual tracking via dense spatio-temporal context learning,” in Computer Vision–ECCV 2014. Springer, 2014, pp. 127–141.

[16] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” 2014.

[17] M. Danelljan, F. S. Khan, M. Felsberg, and J. v. d. Weijer, “Adaptive color attributes for real-time visual tracking,” in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 1090–1097.

[18] Y. Li and J. Zhu, “A scale adaptive kernel correlation filter tracker with feature integration,” in Computer Vision-ECCV 2014 Workshops. Springer, 2014, pp. 254–265.

[19] T. Liu, G. Wang, and Q. Yang, “Real-time part-based visual tracking via adaptive correlation filters,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4902–4912.

[20] Y. Li, J. Zhu, and S. C. Hoi, “Reliable patch trackers: Robust visual tracking by exploiting reliable patches,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 353–361.

[21] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, “Long-term correlation tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5388–5396.

[22] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, “Multistore tracker (muster): A cognitive psychology inspired approach to object tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 749–758.

[23] Chen, Zhe, Zhibin Hong, and Dacheng Tao. "An Experimental Survey on Correlation Filter-based Tracking." arXiv preprint arXiv:1509.05520 (2015).