Huchuan Lu · Dong Wang

Online Visual Tracking

Springer

Online Visual Tracking

Huchuan Lu · Dong Wang

Online Visual Tracking

Springer

Huchuan Lu
Dalian University of Technology
Dalian, China

Dong Wang
Dalian University of Technology
Dalian, China

ISBN 978-981-13-0468-2 ISBN 978-981-13-0469-9 (eBook)
https://doi.org/10.1007/978-981-13-0469-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This book introduces some representative trackers through practical algorithm analysis and experimental evaluations. This book is intended for professionals and researchers who are interested in visual tracking, also students can take it as a reference book. Readers will get comprehensive knowledge of tracking and can learn state-of-the-art methods through this content. In general, the book is organized as follows:

Chapter 1 provides a brief introduction of visual tracking. First, we introduce basic components of tracking including representation scheme, search mechanism, and model update. Then, challenges in visual tracking are displayed. Finally, we show the datasets and evaluation metrics to evaluate trackers in the following chapters.

In Chaps. 2–7, tracking methods based on sparse representation, local model, model fusion, foreground-background segmentation, correlation filters, and advanced deep learning techniques are introduced to understand tracking in a comprehensive view. In each chapter, we first give a brief introduction of the topic and the existing methods. Then, a few representative trackers and their experimental results are discussed in detail. Finally, each chapter ends with a summary.

The last chapter summarizes the book and points out some potential directions of future research for visual tracking.

Dalian, China Huchuan Lu
February 2019

Acknowledgements

First of all, we would like to express our gratitude to our collaborators, Prof. Ming-Hsuan Yang, Mr. Xu Jia, Mr. Wei Zhong, Mr. Dong Wang, Mr. Chong Sun, Ms. Peixia Li, Mr. Boyu Chen, Mr. Lijun Wang, Mr. Pingping Zhang, Mr. Shu Wang, Mr. Fan Yang, Ms. Yunhua Zhang, and Ms. Huilan Jiang. We would also like to thank the organizers of visual tracker benchmark, Dr. Yi Wu, Dr. Jongwoo Lim, and Prof. Ming-Hsuan Yang. Their efforts on the large-scale benchmark make us conduct sufficient experiments for different types of online trackers. We are thankful to the Springer editors, Celine Chang and Jane Li, for their enthusiasm, support, patience, and understanding in the course of our preparation of this work. Part of the content in this book is based on the work supported by the National Natural Science Foundation of China under Nos. 61725202, 61751212, 61829102, 61528101, 61502070, 61472060, and 61272372.

Contents

Chapter 1
Introduction to Visual Tracking

Visual tracking is a rapidly evolving field of computer vision that has been attracting increasing attention in the vision community. One reason is that visual tracking offers many challenges as a scientific problem. Moreover, it is a part of many high-level problems of computer vision, such as motion analysis, event detection, and activity understanding. In this chapter, we give a detailed introduction to visual tracking which includes basic components of tracking algorithms, difficulties in tracking, datasets used to evaluate trackers, and evaluation metrics.

1.1 Basic Components of Visual Tracking Algorithms

Considerable progress in the field of object tracking has been made in the past few decades. Online tracking algorithms typically include three fundamental components, namely, representation scheme, search mechanism, and model update.

Representation Scheme: Object representation is one of the major components of any visual tracking algorithm. Since the early work of Lucas and Kanade (LK) [27], holistic templates (based on raw intensity values) have been widely used for tracking [1, 28]. However, when the visual properties of a target object change significantly, the LK approaches do not perform well. Matthews et al. [28] developed a template update method by exploiting the information of the first frame to correct drifts. Subspace-based tracking approaches have been proposed to effectively account for appearance changes. In [13], Hager and Belhumeur proposed an efficient LK algorithm and used low-dimensional representations for tracking under varying illumination conditions. To enhance tracking robustness, Black and Jepson [6] proposed an algorithm using a pre-trained view-based eigenbasis representation and adopted a robust error norm.

Recently, many tracking methods based on sparse representations have been proposed. For instance, Mei and Ling [29, 30] used a dictionary of holistic intensity templates composed of target and trivial templates. Local sparse representations

H. Lu and D. Wang, *Online Visual Tracking*,
https://doi.org/10.1007/978-981-13-0469-9_1

and collaborative representations for object tracking have also been introduced to handle occlusion. To enhance tracking robustness, a local sparse appearance model was proposed in [26] with the mean shift (MS) algorithm to locate objects. By assuming the representation of particles as jointly sparse, Zhang et al. [45] formulated object tracking as a multi-task sparse learning problem. Zhong et al. [46] proposed a collaborative tracking algorithm that combines a sparsity-based discriminative classifier and a sparsity-based generative model. In [18], sparse codes of local image patches with spatial layout in an object were used to model the object appearance for tracking. To deal with outliers in object tracking, Wang et al. [41] proposed a least soft-threshold squares algorithm by modeling image noise with the Gaussian–Laplacian distribution other than the trivial templates used in [29].

A number of tracking methods based on color histograms have been developed. Comaniciu et al. [9] applied the mean shift algorithm to object tracking on the basis of color histograms. Collins [7] extended the mean shift tracking algorithm to deal with the scale variation of target objects. Birchfield and Rangarajan [5] proposed the spatiogram to capture the statistical properties of pixels and their spatial relationships instead of relying on pixel-wise statistics. A locality sensitive histogram [15] was developed by considering the contribution of local regions at each pixel to clearly describe the visual appearance for object tracking. Histograms of oriented gradients have been adopted for tracking [38] to exploit local directional edge information. Representations based on covariance region descriptors [39] were introduced to object tracking to fuse different types of features. In covariance descriptors, the spatial and statistical properties as well as their correlations are characterized within the same representation. In addition, local binary patterns (LBP) [31] and Haar-like features [40] have been utilized to model object appearance for tracking [4, 11].

Recently, discriminative models have been developed in the field of visual tracking. In these models, a binary classifier is learned online to separate the target from the background. Numerous classifiers have been adopted to visual tracking, and they include support vector machine (SVM), structured output SVM, ranking SVM, boosting, semi-boosting, and online multi-instance boosting. To handle appearance changes, Avidan [2] integrated a trained SVM classifier in an optical flow framework for tracking. In [8], the most discriminative feature combination was learned online to build a confidence map in each frame and thereby separate a target object from the background. In [3], an ensemble of online learned weak classifiers was used to determine whether a pixel belongs to the target region or background. Grabner et al. [11] proposed an online boosting method to select discriminative features for the separation of a foreground object and the background. To balance tracking adaptivity and drifting, Stalder et al. [36] combined multiple supervised and semi-supervised classifiers for tracking. Multiple instance learning (MIL) has also been applied to tracking [4]; that is, all ambiguous positive and negative samples are put into bags to learn a discriminative model. Hare et al. [14] designed a tracking algorithm based on a kernelized structured SVM, which exploits the constraints of predicted outputs.

Several approaches based on multiple representation schemes have also been developed to effectively handle appearance variations. Stenger et al. [37] fused multiple observation models online in a parallel or cascaded manner. Recently, Kwon and Lee [23] developed an object tracking decomposition algorithm that uses multiple observation and motion models to account for relatively large appearance variations caused by drastic lighting changes and fast motion. This approach has been further extended to search for appropriate trackers by Markov chain Monte Carlo sampling [24].

Search Mechanism: Deterministic and stochastic search methods have been developed to estimate object states. When a tracking problem is posed within an optimization framework with an objective function differentiable with respect to motion parameters, gradient descent methods can be used to locate targets efficiently. In [3], the first-order Taylor expansion was used to linearize nonlinear cost functions, and motion parameters were estimated iteratively. Furthermore, mean shift estimation was used to search targets locally by utilizing the Bhattacharyya coefficient as the similarity metric for kernel-regularized color histograms [8]. In [16], Sevilla Lara and Learned Miller proposed a tracking algorithm based on distribution fields; this algorithm allows smoothing objective functions without blurring images and locates targets by searching for the local minimum on the basis of a coarse-to-fine strategy.

However, objective functions for object tracking are usually nonlinear with many local minima. Dense sampling methods have been adopted [4, 11, 14] to alleviate this problem at the expense of a high computational load. Meanwhile, stochastic search algorithms such as particle filters [32] have been widely used because they are relatively insensitive to the local minimum and are computationally efficient. Recent methods based on particle filters have been developed using effective observation models [18, 29, 33] with demonstrated success.

Model Update: The online updating of target representations to account for appearance variations has been known to play an important role in robust object tracking. Matthews et al. [28] addressed the template update problem of the LK algorithm by updating templates using the combination of a fixed reference template extracted from the first frame and the result from the most recent frame. Effective update algorithms have also been proposed in the form of online mixture models [17], online boosting [11], and incremental subspace updating [33].

Considerable attention has been paid to the drawing of samples that are effective in training online classifiers in discriminative models. Grabner et al. [12] formulated the update problem as a semi-supervised task in which the classifier is updated with labeled and unlabeled data; such task is in contrast to supervised discriminative object tracking. To handle ambiguously labeled positive and negative samples obtained online, Babenko et al. [4] focused on the tracking problem within the multiple instance learning framework and developed an online algorithm. To exploit the underlying structure of unlabeled data, Kalal et al. [19] developed a tracking algorithm within the semi-supervised learning framework to select positive and negative samples for model updating. In [14], a tracking algorithm that directly predicts target location changes between frames on the basis of structured learning was proposed. Yu et al. [44] presented a tracking method that is based on co-training to combine generative and discriminative models. Although considerable progress has clearly been made, developing an adaptive appearance model without drifts remains to be difficult.

1.2 Challenges in Visual Tracking

Many difficulties plague visual tracking, and they include occlusion, scale variation, deformation, fast motion, motion blur, background clutter, in-plane rotation, out-of-plane rotation, illumination variation, out-of-view, and low resolution. The descriptions of these challenges are shown in Table 1.1, and some visual examples are illustrated in Fig. 1.1.

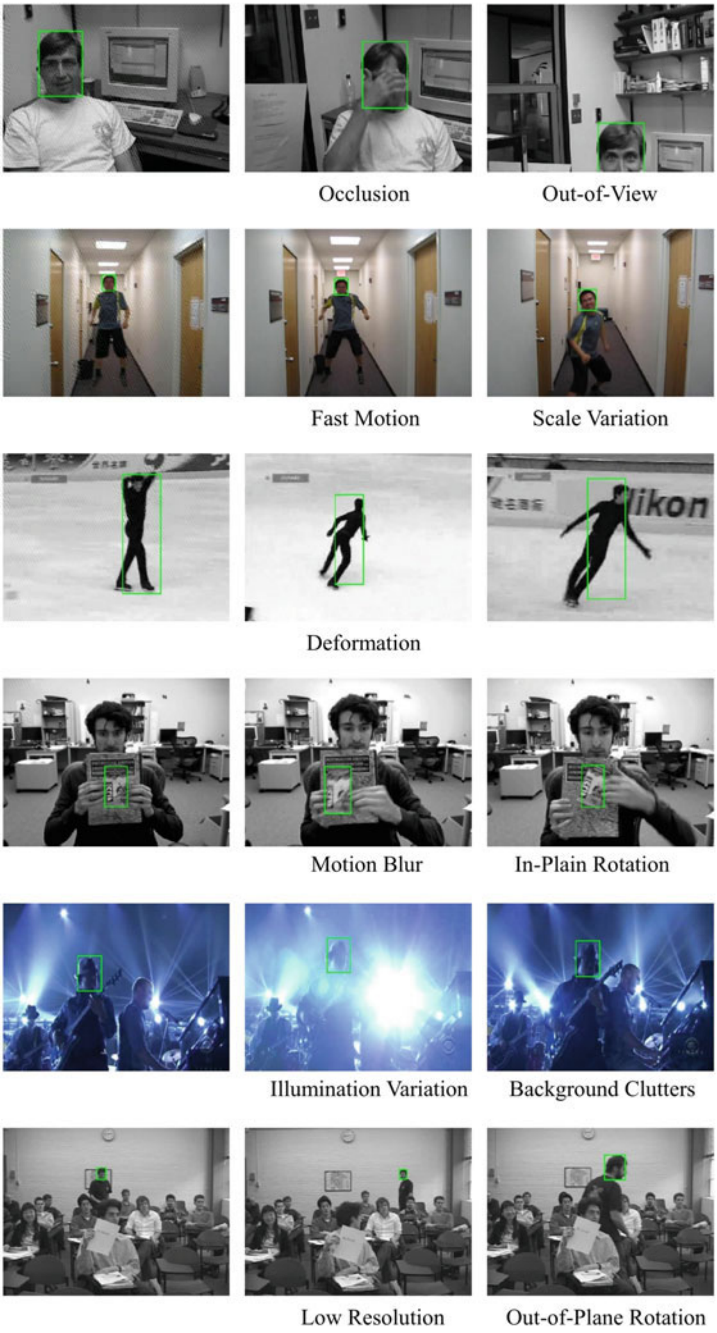
1.3 Tracking Datasets

In this section, we briefly introduce the commonly used datasets for evaluating the performance of different online trackers.

OTB: OTB-2013 [42] contains 50 target objects and OTB-2015 [43] expands the sequences in OTB-2013 to include 100 target objects in the tracking benchmark TB-100 dataset. As some of the target objects are similar or less challenging, 50 difficult and representative ones in the TB-50 dataset were also selected for an in-depth analysis. Note that as humans are the most important target objects in practice, the

Table 1.1 Description of challenging factors in visual tracking

Challenge	Description
Occlusion	The target is partially or fully occluded
Scale variation	The ratio of the bounding boxes of the first frame and the current frame is out of range
Deformation	The object has nonrigid deformation
Fast motion	The motion of the ground truth is too large
Motion blur	The target region is blurred due to the motion of the target or the camera
Background clutter	The background near the target has a similar color or texture as the target
In-plane rotation	The target rotates in the image plane
Out-of-plane rotation	The target rotates out of the image plane
Illumination variation	The illumination in the target region is significantly changed
Out-of-view	Some portion of the target leaves the view
Low resolution	The number of pixels inside the ground truth bounding box is too small



Occlusion

Out-of-View

Fast Motion

Scale Variation

Deformation

Motion Blur

In-Plain Rotation

Illumination Variation

Background Clutters

Low Resolution

Out-of-Plane Rotation

Fig. 1.1 Some visual examples of different challenges in visual tracking

TB-100 dataset contains more sequences of this category (36 body and 26 face/head videos) than of others.

For a satisfactory analysis of the strengths and weaknesses of tracking algorithms, TB-100 categorizes sequences according to 11 attributes, namely, illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutter, and low resolution. Each attribute represents a specific challenging factor in object tracking. One sequence may be annotated with many attributes, and some attributes occur more frequently than others do. The characteristics of tracking algorithms can be clearly analyzed from the sequences with the same attributes. For example, to evaluate how well the tracker handles occlusion, one may use 49 sequences (29 in TB-50) annotated with the OCC attribute.

VOT: The VOT challenges provide the visual tracking community with a precisely defined and repeatable way of comparing short-term trackers as well as a common platform for discussing the evaluation and advancements that are made in the field of visual tracking.

The VOT2015 dataset consists of 60 short sequences annotated with 6 different attributes, namely, occlusion, illumination change, motion change, size change, camera motion, and unassigned. The major difference between VOT2015 and OTB-2015 is that the VOT2015 challenge provides a re-initialization protocol (i.e., trackers are reset with ground truth in the middle of the evaluation if tracking failures are observed).

In [20], the VOT committee analyzed the properties of an average overlap with and without resets in terms of tracking accuracy estimator. The analysis showed that measures with resets can drastically reduce bias. More important, the no-reset measure becomes reliable only on extremely large datasets. Hence, large variances of no-reset estimators combined with small numbers of sequences can distort performance measurements. As datasets typically do not contain sequences of equal lengths, variances are even increased. VOT2013 [22] introduced a ranking-based methodology that accounts for the statistical significance of results, and this methodology was extended in VOT2014 with tests of practical differences [21]. VOT2015 follows the VOT2014 challenge and considers the same class of trackers.

TColor-128 [25]: TColor-128 is a large dataset with 128 color sequences and is devoted to color visual tracking. Sequences in TColor-128 mainly come from new collections and previous studies, such as OTB-2013 and VOT2013. The new collections of TColor-128 contain 78 color sequences newly collected from the Internet. The 78 sequences largely increase the diversity and difficulty of the previous 50 sequences as they involve various circumstances, such as highways, airport terminals, railway stations, concerts, and so on. Similar to that in the OTB dataset [42, 43], each sequence in TColor-128 is also annotated by its challenge factors with 11 attributes.

PTB [35]: PTB has 100 video clips with RGB and depth data. These videos are captured by a standard Microsoft Kinect 1.0, which uses a paired infrared projector