Synthesis Lectures on Computer Vision

SYNTHESIS
COLLECTION OF TECHNOLOGY

Ning Xu · Weiyao Lin ·
Xiankai Lu · Yunchao Wei

Video Object Tracking

Tasks, Datasets, and Methods

Springer

Synthesis Lectures on Computer Vision

This series publishes on topics pertaining to computer vision and pattern recognition. The scope follows the purview of premier computer science conferences, and includes the science of scene reconstruction, event detection, video tracking, object recognition, 3D pose estimation, learning, indexing, motion estimation, and image restoration. As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner. As a technological discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems, such as those in self-driving cars/navigation systems, medical image analysis, and industrial robots.

Ning Xu · Weiyao Lin · Xiankai Lu ·
Yunchao Wei

Video Object Tracking

Tasks, Datasets, and Methods

Springer

Ning Xu
Adobe Research
San Jose, CA, USA

Xiankai Lu
School of Software
Shandong University
Jinan City, China

Weiyao Lin
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China

Yunchao Wei
School of Computer and Information
Technology
Beijing Jiaotong University
Beijing, China

ISSN 2153-1056 ISSN 2153-1064 (electronic)
Synthesis Lectures on Computer Vision
ISBN 978-3-031-44659-7 ISBN 978-3-031-44660-3 (eBook)
https://doi.org/10.1007/978-3-031-44660-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

With the universal access of cameras in all kinds of devices (e.g., mobile phones, surveillance cameras, and in-vehicle cameras), video data has gone through an exponential increase nowadays. For a lot of video-related applications such as autonomous driving, video editing, and augmented reality, segmenting target objects is important to help understand video content.

Video object tracking (VOT) is a fundamental task for video understanding in computer vision. VOT can be divided into different settings further, which are slightly different in the definition and focus areas. In this book, we will give an throughout introduction of the task of VOT (including different settings). In Chap. 1, we first briefly introduce the settings of VOT. Then in Chap. 2, we introduce the VOT task under three most common problem settings which are single-object tracking, multi-object tracking, and multi-object tracking and segmentation. Finally, in Chap. 3, we give introduction to our held HiEve challenge on VOT.

USA
September 2022

Ning Xu

Contents

Introduction

1

The ubiquity of cameras in various devices (e.g., mobile phones, surveillance cameras, in-vehicle cameras) has led to an exponential increase in video data. For a lot of video-related applications, being able to understand, localize, and track target objects is one of the most fundamental and important problems. For example, in the field of autonomous driving, precise localization of surrounding cars and pedestrians is crucial for vehicles to avoid collisions and take safe actions. Additionally, in the field of augmented reality (AR), to attach AR effects to some moving object in the space and to make the effect look realistic, we also need some algorithm to follow the motion of the target object in real-time.

Video object tracking (VOT), a fundamental problem for video understanding in computer vision, is the underlying technology that enables the above applications. In a formal definition, the task of VOT aims at producing tight bounding boxes around one or multiple target objects in the video. Meanwhile, VOT focuses on the ability to track target objects over a long period of time in videos. However, it will face big challenges when objects have fast motion, large appearance changes, similar instances, and heavy occlusion etc.

This book provides a comprehensive introduction to VOT. Specifically, in Chap. 2, we introduce the VOT task under three most common problem settings: single-object tracking, multi-object tracking, and multi-object tracking and segmentation. Each problem setting is first elaborated with a formal problem definition, followed by presenting well-known challenges, the most popular datasets and evaluation metrics. Then we overview different types of methods that have been proposed for the problem. Finally, we select two or three most representative and effective methods and dive deep into their idea details as well as experiments. Additionally, we also include the results of the recent impactful competition (HiEve Challenge) on VOT in Chap. 3, offering readers the methods and results of top-performing teams.

N. Xu et al., *Video Object Tracking*, Synthesis Lectures on Computer Vision,
https://doi.org/10.1007/978-3-031-44660-3_1

Table 1.1 Comparison between VOT tasks on several attributes

Task	Classification	Tracking	Detection	Segmentation	Object num
VOT					
SOT	×	√	×	×	Single
MOT	×	√	√	×	Multiple
MOTS	×	√	√	√	Multiple

Before diving into each individual task in the following chapters, we first give an overview of each task and let readers understand the goal of each task and the difference among them.

- Single object tracking (SOT) is a VOT setting where the target's bounding box in the first frame is given and algorithms need to predict the target position in the subsequent frames.
- Multi-object tracking (MOT) can be easily distinguished from SOT because it requires to extract the spatial and temporal trajectories of multiple moving objects from the video.
- Multi-object tracking and segmentation (MOTS) is a recently proposed tracking task that requires detecting, tracking, and segmenting objects belonging to a set of given classes from the video.

Besides the above differences, in Table 1.1 we list other attributes that can further understand and distinguish these tasks. First, all these VOT tasks do not require addressing the classification problem (i.e., recognizing object categories). Second, each task involves a different combination of sub-problems. For instance, the SOT and MOT do not need to solve the segmentation problem while the MOTS need. Lastly, the two tasks MOT and MOTS need to handle multiple instances in the video, while only the SOT focuses on a single target object.

Tracking

2

In this chapter, we will elaborate on the task of video object tracking (VOT), which aims at producing tight bounding boxes around one or multiple target objects in the video. The VOT task has different problem settings given different input or output requirements. Single object tracking (SOT) Sect. 2.1 gives the bounding box of object that needs to track in the first frame and requires to predict the whole tracklet of this object. Multi-object tracking (MOT) Sect. 2.2 requires to detect the objects first and then get their tracklets. Multi-object tracking and segmentation (MOTS) Sect. 2.3 need to predict segmentation further.

2.1 Single Object Tracking

2.1.1 Introduction

In this section, we introduce single object tracking, that given the target's bounding box in the first frame and predicting the target position in the subsequent frames. Video object tracking is one of the important tasks in computer vision, and has a wide range of applications in real life, such as video surveillance, visual navigation, etc. Video target tracking tasks also face many challenges, such as target occlusion, target deformation and so on. In order to solve the challenges in target tracking and achieve accurate and efficient target tracking, a large number of target tracking algorithms have appeared in recent years.

This section introduces the basic principles, improvement strategies and representative work of the two mainstream algorithm frameworks in the field of video target tracking (target tracking algorithm based on correlation filtering and Siamese network) in the past ten years. The target tracking algorithm also introduces typical solutions to various problems from the perspective of solving the challenges faced by target tracking, and summarizes the historical development and future development trend of video target tracking. This section also introduces and compares the datasets and challenges for object tracking tasks in

N. Xu et al., *Video Object Tracking*, Synthesis Lectures on Computer Vision,
https://doi.org/10.1007/978-3-031-44660-3_2

detail, and summarizes the characteristics and advantages of various video object tracking algorithms based on the data statistics of the datasets and the evaluation results of the algorithms.

SOT has been widely studied with the development of computer vision history. First, this setting is easy and flexible for many real-world applications, such as human trajectories construction in public security surveillance systems, navigation in Autonomous driving and unmanned aerial vehicles, Motion Trajectory Capture and Active Tracking in Robotics and pose tracking in human-object-interaction. In the late 90s, *generative methods* are the mainstream solution for SOT [166]. These methods are under the geometry-based methods with the strong assumption, such as optical flow (TLD) [85, 90], Kalman Filter [153], partial filter [3], Meanshift [46] with hand-craft feature, such as SIFT and SURF.

From the 2000s, *learning based* SOT methods become the mainstream [193] with the development of machine learning. These methods can be categorised into two classes: shallow learning and deep learning. Shallow learning ones combine the classical machine learning algorithms, such as SVM [68, 184], ensemble learning [7, 231], sparse coding [223] and correlation filter [75]. Tian et al. [184] considered tracking as a binary classification problem, and SVM is selected as the main classification. In object tracking, the target area defined as positive data, and the surrounding environment is defined as negative data. The goal is to train an SVM classifier that can classify positive and negative data into new frames.

In the deep learning era, early methods explored fully convolution network or deep correlation filter for target prediction [134, 172, 173, 194]. The tracking pipeline still follow into the traditional tracking paradigm by using the first frame annotation to generate many samples for network finetuning. Then, Siamese architectures [15, 54, 109, 110, 183, 227] or variants of recurrent neural networks [67] becomes the popular tracking methods. These methods tend to train a fully-convolution siamese networks with annotated videos and apply the fixed weight model to the unseen video directly [179]. With the development of graph neural network and Transformer, some researchers have apply these powerful relation modeling method for SOT, for instance, STARK [212], Keeptrack [139], TransT [26] and CSWinTT [175].

In the rest of this section, we will first introduce the common challenges of this task Sect. 2.1.2, popular benchmarks and evaluation metrics Sect. 2.1.3 as well as overview of common and effective methods Sect. 2.1.4. Then we present three representative SOT methods in detail which belong to Correlation filter methods Sect. 2.1.5, one-stage deep tracking methods Sect. 2.1.6, Siamese network methods Sect. 2.1.7, respectively.

2.1.2 Challenges

In this section we discuss a few challenging problems in SOT. As a classical task in the computer vision, SOT faces lots of challenges:

- *Appearance variation*. Appearance change is a common interference problem in target tracking. When the appearance of a moving target changes, its characteristics and appearance model will change, which may easily lead to tracking failure. For example: athletes in sports competitions, pedestrians on the road due to the fast motion and view changes.
- *Scale variation*. Scale adaptation is also a key issue in target tracking. When the target scale is reduced, since the tracking frame cannot be adaptively tracked, a lot of background information will be included, resulting in an update error of the target model: when the target scale When it increases, since the tracking frame cannot completely include the target, and the target information in the tracking frame is incomplete, the update error of the target model will also be caused. Therefore, it is very necessary to achieve scale adaptive tracking.
- *Occlusion and disappearance*. The target may be occluded or disappear briefly during the movement. When this happens, the tracking frame will easily include the occlusion clutter and background information in the tracking frame, which will lead to tracking in subsequent frames. The target drifts to the clutter. If the target is completely occluded, the tracking will fail because the corresponding model of the target cannot be found.
- *Other factors* Motion blurring Changes in light intensity, fast movement of the target, low resolution, etc. will lead to image models, especially when the moving target is similar to the background. Therefore, select effective features to distinguish the target from the background. Very necessary.

Based on these discussion, we can see that robust SOT method need to satisfy the following criteria: the learned feature representation should be variance for different scenarios, especially for occlusion case meanwhile be adaptive for appearance change. We will give more detailed explanation in the method sections.

2.1.3 Dataset and Metric

This section will detail popular public dataset as well as common evaluation metrics. Considering different tracking datsets have different evaluation metrics. Therefore, we introduce each dataset with corresponding evaluation metrics.

OTB (object tracking benchmark) is a old and classical dataset [207] which contains two version: OTB100 (100 videos) and OTB50 (50 videos). The videos are labeled with 11 attributes, namely: illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, inplane rotation, out-of-plane rotation, out-of-view, background clutters, and low resolution. There are two widely used metrics for OTB datasets: distance precision and overlap success that are computed per-frame and across video on average:

$$P^t = \|C_{tr}^t - C_{gt}^t\|_2 \quad S^t = \frac{BB_{tr} \cap BB_{ft}}{BB_{tr} \cup BB_{ft}} \tag{2.1}$$