J. Vis. Commun. Image R. 34 (2016) 37–49

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

A novel hierarchical data association with dynamic viewpoint model for multiple targets tracking ☆

Yang Yi [a,c,*], Zengwei Mo [b], Jie-wen Tan [d]

[a] School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[b] School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China
[c] SYSU-CMU Shunde International Joint Research Institute, Shunde, Guangdong, China
[d] Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 19 January 2015
Accepted 14 October 2015
Available online 21 October 2015

Keywords:
Online multiple targets tracking
Tracking-by-detection
Data association
Dynamic viewpoint model
Structural context model
Branch partition
Candidate upgrading
Incremental motion pairing inference

ABSTRACT

A new framework of hierarchical data association tracking (HDAT) with branch partition, candidate upgrading and incremental motion pairing inference is proposed to resolve the problem of online multiple targets tracking. Branch partition divides the process into several independent parts so as to reduce the computational complexity on affinity. Candidate upgrading improves the robustness of target initialization by tracking potential targets and incremental motion pairing inference could benefit the occlusion handling. Furthermore, a dynamic viewpoint model (DVM) and its iterative computation algorithm are developed for tracking multiple targets under moving camera videos. Extensive data experiments on several public benchmarks show that the presented approach achieves comparable results to state-of-the-art on static camera videos and promising results on moving camera videos, and moreover, the runtime performance is significantly improved.

1. Introduction

It has been noticed that multiple targets tracking is becoming more and more important in computer vision and image processing applications, such as intelligent surveillance, human–computer interaction and automatic driving [1–3,48,49]. This paper addresses the problem of simultaneously tracking a variable number of targets with specified categories from a single camera, which has been believed challenging since there are many uncertain factors such as measurement noise, cluttered background, occlusions, the changeable number of targets, and varying targets' appearances and motions [4–8,50].

With the development of theories and technologies of object detection, tracking-by-detection framework (e.g. [4–6]) has become one of the most impactful methods to deal with above difficulties, although, even the advanced detectors still yield false alarms and missed detections [9]. It's suggested that introducing spatio-temporal constraints and structural context information into the tracking-by-detection framework may improve the performance of predicting target motion, completing target trajectory, reducing ID switches, and trajectory fragmentations.

To exploit the spatio-temporal information, tracking-by-detection could be categorized broadly into global optimization tracking [10,11] and online tracking [4,51]. The former one defers the association decision and optimizes multiple trajectories simultaneously in the whole video or within a temporal window [6,12]. By using the future frames, it may overcome some of the detection failures and enhance robustness. However, its heavy computational complexity and decision latency limits its applications to offline or the areas without critical requirement on time (time-critical applications) [9,13]. Online tracking only utilizes the information up to the current frame and could deal with frame-by-frame data association, thus it could be applied to real-time applications. For online multiple targets tracking, it's very difficult to cope with inaccurate (or even missed) detections of occluded targets [14], and targets tend to drift after long time occlusions [9,15].

The tracking performance on time-critical applications does not merely depend on accuracy, but also on time performance. In this paper, a new hierarchical data association tracking (HDAT) framework belonging to online tracking is proposed to improve the overall tracking performance.

In most of the time, data association in each frame can be divided into smaller independent sub-problems to reduce

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author at: School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China.

E-mail address: issyy@mail.sysu.edu.cn (Y. Yi).

http://dx.doi.org/10.1016/j.jvcir.2015.10.010
1047-3203/

unnecessary calculation. Motivated by the idea of dividing and conquering, this paper develops branch partition to make data association perform independently within each branch. Thus the efficiency of the presented method is improved by reducing affinity calculation during the data association.

In order to overcome the false alarms and avoid incorrect target initialization, online multiple targets tracking usually accumulates evidences for target initialization, for example, taking consecutive overlapping detections as the target initialization strategy [4]. In the paper, the process of target initialization is integrated into the framework of HDAT, and the association between the candidates and the detections could be similar to the process between the detections and the targets. An independent affinity model to enrich the accumulation of initialization evidences is developed for the association of detections–candidates, and meanwhile new candidate upgrading strategies are presented to make the whole target initialization more flexible and quantifiable.

The structural context has been widely applied in object detection and has become increasingly important in multiple targets tracking for occlusion reasoning, miss detections handling and so on. In pedestrians tracking, some current researches (e.g. [2,16–18]) show that pedestrians tend to walk in groups, and pedestrians within the same group are likely to follow a similar motion pattern, which can be a valuable structural context cue for pedestrian tracking. The online tracking cannot access the targets' full trajectories to infer grouping information. Therefore an incremental motion pairing inference is presented to infer the grouping relationship between the targets online. The grouping information can help to update a target's motion state when the target fails to be associated with any detection.

The coordinate mapping from 3D real world to 2D image plane during imaging process may introduce some noises, thus obtaining the coordinate mapping will be helpful for the following noise removal. To obtain more accurate mapping, Hoiem et al. [19] proposed a single-frame viewpoint model to model the scene and the relationships between targets and environments, so as to improve object detection performance and provide scene parsing information. Inspired by [19], this paper proposes a dynamic viewpoint model (DVM) and integrates it into HDAT. And then HDAT with DVM (HDAT-DVM) is used to overcome the difficulties on moving camera videos. By pruning out false detections and targets, DVM parses specific scene efficiently in order to get an approximate 3D coordinate mapping and improve the tracking performance.

The main contributions of our works are as follows. (1) A new framework to deal with the problem of online multiple targets tracking under a single camera is proposed, which including of branch partition, candidate upgrading, and incremental motion pairing inference strategies. (2) A dynamic viewpoint model and its iterative computation algorithm (ICA) are developed to supply more scene parsing information and improve the tracking performance by filtering out false detections and targets.

The rest of this paper is organized as follows: Section 2 discusses the related work, and Section 3 gives a detailed description of the presented HDAT framework. Section 4 introduces DVM and its iterative computation algorithm. The experimental results and analysis are presented in Section 5, followed by the conclusions in Section 6.

2. Related work

2.1. Tracking-by-detection

Global optimization usually optimizes detection assignments over large temporal windows [6,10]. The optimal framework includes dynamic programming [20], network flows [21], maximum weight independent set [22], generalized minimum clique graphs [23], k-shortest paths [24], etc. These methods operate on a large sliding window and require detections of several consecutive frames as input, which precludes their online applications [25]. Online tracking [4,9,14,15] builds trajectories sequentially based on frame-by-frame association and can be formulated as an assignment problem. However, the lack of information from future frames makes it easy to produce ID switches and target drifts [10]. In order to overcome the difficulties, Breitenstein et al. [4] merged both the intermediate output of the detector (i.e. detector confidence) and the output of online-learned classifiers into the particle filter framework to compute the weight of a particle. However, [4] needed to utilize target-specific classifiers to help handling missing detections. Possegger et al. [9] exploited geometric information such as occlusion cues, detector reliability and target motion prediction to tackle detection failures. Wu et al. [51] proposed a motion agreement method to select stable object regions to track dynamically. Yan et al. [15] optimized the association between the tracked targets and the output of independent trackers and the detector. Different from the hierarchical data association steps in [15], the presented framework supplies new level design, affinity model and target initialization strategy.

2.2. Structural context model

Structural context describes the diversity of the relationships, such as among the parts of one target [5], between the targets and the scene [26], or among the targets [27,47]. Many methods [2,45] showed that social grouping information is an important structural context in multiple targets tracking. Yet, the relationship between target tracking and group inferring is a chicken-and-egg problem: knowing group relations is helpful to estimate targets' trajectories, but without additional information, it is very hard to discover possible groupings if there are no precise target trajectories [28]. Recently, most of the solutions combined group inference and trajectories/motion estimation together. Yan et al. [45] modeled the target motion behavior from the first-person perspective and exploited it to track target. Yan et al. [46] presented a hierarchical group structure and optimized it. Yamaguchi et al. [29] modeled pedestrian behaviors (e.g. speed, direction, attraction) explicitly and integrated them into an energy function. Unfortunately, the computation cost of the method may be a little bit high. Qin and Shelton [2] introduced the concept of group mean trajectory and measured the distance between individual tracklet and group mean trajectory, then added the distance to the original objective function. However, this method belongs to the family of global optimization. Without using any complicated social grouping model which may bring about lots of computational complexity, the presented method infers the frame-wise pairing relationship online and utilizes the pairing relationship instead of obtaining the targets' full trajectories by offline method.

2.3. Tracking on moving camera video

The tracking task is believed to be challenging when camera moves. To address the difficulties of detection and tracking under moving camera, several approaches [25,30] have been developed. Wojek et al. [30] proposed a probabilistic 3D scene model to track multiple targets. Ess et al. [25] presented a graphical model to solve pedestrian detection and tracking, which, however, needs a stereo camera setup. Hoiem et al. [19] combined surface geometry and camera viewpoint together to improve the performance of object detection, and they only adopted the approaches in single image processing. Unlike [19], DVM takes temporal information of video into consideration and is used to filter out the implausible detections and targets. With the aim of being applied to robot nav-