Computer Vision and Image Understanding 117 (2013) 1257–1272

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Multi-target tracking on confidence maps: An application to people tracking

Fabio Poiesi *, Riccardo Mazzon, Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London, London, UK

ARTICLE INFO

Article history:
Received 16 March 2012
Accepted 23 August 2012
Available online 29 November 2012

Keywords:
Track-before-detect
Crowd
Multi-target tracking
Markov Random Field
Gaussian Mixture Model
Likelihood modeling

ABSTRACT

We propose a generic online multi-target *track-before-detect* (MT-TBD) that is applicable on confidence maps used as observations. The proposed tracker is based on particle filtering and automatically initializes tracks. The main novelty is the inclusion of the target ID in the particle state, enabling the algorithm to deal with unknown and large number of targets. To overcome the problem of mixing IDs of targets close to each other, we propose a probabilistic model of target birth and death based on a Markov Random Field (MRF) applied to the particle IDs. Each particle ID is managed using the information carried by neighboring particles. The assignment of the IDs to the targets is performed using Mean-Shift clustering and supported by a Gaussian Mixture Model. We also show that the computational complexity of MT-TBD is proportional only to the number of particles. To compare our method with recent state-of-the-art works, we include a postprocessing stage suited for multi-person tracking. We validate the method on real-world and crowded scenarios, and demonstrate its robustness in scenes presenting different perspective views and targets very close to each other.

1. Introduction

Multi-target tracking is a challenging task in real-world scenarios due to the variability of target movements, shapes, clutter and occlusions. Moreover, the computational cost may exponentially increase with the number of co-occurring targets and the maximum number of targets may have to be fixed *a priori*. Single-target tracking generally represents the state of each target with a single state vector [1]. In multi-target tracking the size of state vector increases with the number of targets [2–12] unless a single-target tracker is initialized for each target [13–20]. We refer to the former approach as *one-state-per-target* (OSPT) and to the latter *one-filter-per-target* (OFPT). OSPT methods perform tracking optimization at each time step on the overall state space. Only a predefined number of targets can be tracked [14] or ad hoc stages can be used to estimate the number of targets in the scene [2,5]. OFPT methods perform tracking by a local optimization for each target, thus limiting their application to situations with a small number of targets that are easily distinguishable.

Target locations may be gathered from sensors (e.g. laser, sonar, camera) via confidence maps that provide multiple measurements per target and carry information in the form of intensity levels over space (Fig. 1). These intensity levels are affected by different types of noise on background areas and/or on the targets themselves, thus resulting in inaccurate position estimations. Tracking algorithms employ target locations as measurements, either directly as confidence maps (unthresholded data) [13,21,20,22] or as binary maps (target/non-target information) obtained by thresholding the confidence values [3,4,6,10]. Although the latter strategy is the most commonly used, relevant data may be lost with this process. Tracking-by-detection methods [20] perform target-tracker association, and initialization and termination of tracks with greedy algorithms. Track-before-detect (TBD) methods perform tracking of targets using unthresholded data [23] and target-tracker association is implicitly computed by the tracker. TBD is a Bayesian filter, generally built on the concept of particle filter, and commonly used for radar tracking [23,24]. Multi-target tracking is performed on noisy intensity levels and the targets are assumed to be point targets. Initialization and termination of tracks are performed by the tracker using target *birth and death* models.

In this paper we propose a novel multi-target tracker based on TBD algorithm [23] and applied to confidence maps. To enable multi-target tracking, we develop a method where target IDs are assigned based on Mean-Shift clustering and Gaussian Mixture Model (GMM). The birth and death of targets are modeled with a Markov Random Field (MRF). Unlike [24], we do not need to define the maximum number of targets *a priori* and, unlike [20], the initialization of a track may occur in any location of the image, thus making the multi-target *track-before-detect* (MT-TBD) automatic and flexible to different scenarios. MRF enables multi-target tracking without augmenting the state (OSPT methods, e.g. [2]) or the number of filters (OFPT methods, e.g. [13]), caused by an increase

* Corresponding author.
E-mail address: fabio.poiesi@eecs.qmul.ac.uk (F. Poiesi).
URL: http://www.eecs.qmul.ac.uk/~andrea/ (A. Cavallaro).

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2012.08.008

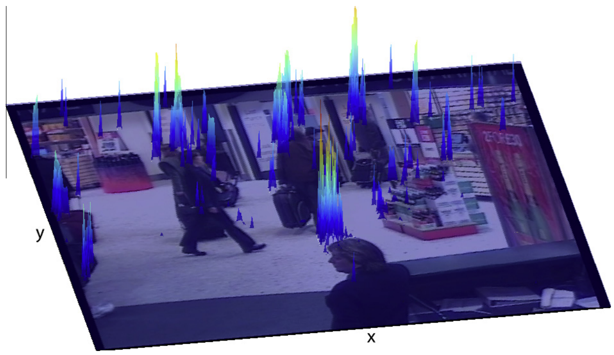

Fig. 1. Sample confidence map that we use as input (observation) to simultaneously track multiple objects. In this example, the confidence map is obtained with a head localization method based on [25].

in the number of targets. Moreover, the use of MRF overcomes the limitations of Buzzi et al. [24] by allowing a reliable tracking of close targets without loss of performance and leads to a computational complexity depending only on the number of particles. Compared to the recent work by Benfold and Reid [10], the tracking accuracy of the proposed MT-TBD improves by 11% with 2 s of latency and by 10% with 4 s of latency on a publicly dataset from Oxford town center.

The paper is organized as follows. Section 2 discusses the related work for multi-person tracking. Section 3 gives an overview of the proposed approach and introduces MT-TBD. The ID management via MRF is explained in Section 4. Section 5 illustrates the application of MT-TBD to multi-person tracking. Section 6 discusses the experimental results, the comparisons with existing methods and the analysis of the computational complexity. Finally, in Section 7 we draw the conclusions and present possible research directions.

2. Related work

In this section we discuss recent works on multi-person tracking, we analyze their main contributions and classify each method in its corresponding category. Multi-target video trackers can be classified into causal and non-causal methods. *Causal* methods use information from past and present observations to estimate trajectories at the current time step. *Non-causal* methods use also information from future observations, thus resulting in a delayed decision. Although non-causal approaches are not suitable for time-critical applications, they can achieve a global optimum leading to more robust results during occlusions.

Examples of causal trackers are Bayesian filters [17,10,16,15,20]. Yang et al. [17] use a Bayesian-based detection association obtained by Convolutional Neural Network (CNN) trained on color histograms, elliptical head model, and bags of SIFTs. Benfold and Reid [10] find the optimum trajectories within a four-second window by a Minimum Description Length (MDL) method applied on trajectories from a forward and backward Kanade–Lucas–Tomasi (KLT) tracking and from a Markov Chain Monte Carlo Data Association (MCMCDA). Alternatively, the particle filter is used in [16,15,20]. Ali and Dailey [16] track heads obtained by Haar-like features and AdaBoost; whereas Xing et al. [15] employ the Hungarian algorithm for the optimization of short but reliable trajectories obtained by tracking the upper human body. Depending on the scenario, Breitenstein et al. [20] track people detected by Histogram of Oriented Gradients (HOG) or Implicit Shape Model (ISM). Here the association between detections and tracks is performed by a greedy algorithm and boosting. A different approach is presented by Rodriguez et al. [7] where tracking is obtained on four points per head by KLT and head detection is optimized by crowd density estimation and camera-scene geometry. Tag-and-track methods for high-density crowd are proposed in [26,27], where targets are assumed to follow a learned crowd behavior. Ali and Shah [26] deal with crowds with coherent motion by modeling their global behavior, the environment structure and the local behavior of people. Rodriguez et al. [27] focus on crowds with non-coherent motion where the modeling is performed by Correlated Topic Model (CTM) that predicts the next position of a person by exploiting the optical flow. Note that among causal methods, only Benfold and Reid [10] and Rodriguez et al. [7] use an OSPT framework. This is because the OSPT is generally more complex than OFTP, but the modeling for multi-person tracking is more flexible and computationally cheaper [10].

As for non-causal trackers, short-term tracks (tracklets) [3,4,8,6,9,11,12] can be associated over time by using a modification of the Multi-Hypothesis Tracking (MHT) algorithm [28], where the detections are obtained with a person detector [29]. Huang et al. [3] associate tracklets by Hungarian algorithm using position, time and appearance features, and then refine them using entry and exit points in the scenes, which are in turn learned from tracklets. Li et al. [4] show how the association can be improved by using a combination of RankBoost and AdaBoost in a hierarchical approach where longer trajectories are generated using a set of 14 features per tracklet by starting from the lower levels. In Yang et al. [8], the association is performed using RankBoost applied to an optimization of affinities and dependencies between tracklets by a Conditional Random Field (CRF). Kuo et al. [6] associate tracklets using an AdaBoost classifier that learns online the discriminative appearance of targets based on their color histogram, covariance matrix features and HOG. Kuo et al. [9] extract motion, time and appearance from different body parts of each target in order to perform a re-identification step to resolve long-term occlusions. Yang and Nevatia [11] learn online the non-linear motion of people and a Multiple Instance Learning (MIL) framework for the appearance modeling using the estimation of entry and exit regions. Furthermore, Yang and Nevatia [12] use CRF to model affinity relationships between tracklet pairs, where the association of tracklets is based on Hungarian algorithm and a heuristic search. Table 1 summarizes the methods covered in this section and the dataset on which these methods have been tested.

Similarly to Stalder et al. [21] and Breitenstein et al. [20], the proposed MT-TBD is a causal method that makes use of confidence maps as measurement for tracking. However, compared to [21], we use the confidence maps online without the need of any temporal processing and, compared to [20], an automatic assignment between confidence map and targets is performed. Moreover, unlike [20], which uses manually selected areas at the borders of the image to initialize tracks, we do not use any prior information about the scene. This becomes extremely advantageous when targets temporarily undergo a total occlusion in any position of the image. In addition to this, we overcome the limitations of OFTP approaches [20,22] with a global and instantaneous optimization of target tracking in MT-TBD by employing a general likelihood function obtained from a controlled sequence (Section 5.1). Finally, unlike De Leat et al. [22], the use of multiple measurements per target is tested in various crowded scenes with different camera perspectives.

3. Sequential Monte Carlo estimation for multi-target track-before-detect

3.1. Confidence maps and track-before-detect

Let a confidence map $\mathfrak{M}$ provide the information on the estimated position of targets through spatially-localized intensity levels (Fig. 1). The ideal representation of the target position on a