Computer Vision and Image Understanding 154 (2017) 166–181

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Combined segmentation, reconstruction, and tracking of multiple targets in multi-view video sequences

M. Babaee*, Y. You, G. Rigoll

Institute for Human-Machine Communication, Technical Univ. of Munich, Munich, Germany

ARTICLE INFO

Article history:
Received 17 May 2016
Revised 7 August 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
Superpixels
Segmentation
Reconstruction
Tracking
Hypergraph

ABSTRACT

Tracking of multiple targets in a crowded environment using tracking by detection algorithms has been investigated thoroughly. Although these techniques are quite successful, they suffer from the loss of much detailed information about targets in detection boxes, which is highly desirable in many applications like activity recognition. To address this problem, we propose an approach that tracks superpixels instead of detection boxes in multi-view video sequences. Specifically, we first extract superpixels from detection boxes and then associate them within each detection box, over several views and time steps that lead to a combined segmentation, reconstruction, and tracking of superpixels. We construct a flow graph and incorporate both visual and geometric cues in a global optimization framework to minimize its cost. Hence, we simultaneously achieve segmentation, reconstruction and tracking of targets in video. Experimental results confirm that the proposed approach outperforms state-of-the-art techniques for tracking while achieving comparable results in segmentation.

1. Introduction

Tracking of multiple targets in a crowded and unconstrained environment has many applications in video surveillance and security systems. This is a challenging problem due to the high amount of noise in the measured data, occlusion among targets, and interaction of targets with themselves or with other objects. Currently, tracking-by-detection is considered as the most successful solution for this problem Ben Shitrit et al. (2011); Milan et al. (2014); 2013); Park et al. (2015); Zamir et al. (2012). However, tracking of detection boxes is not enough for many real applications such as human activity recognition and analysis.

In this work, we propose an approach to track segmented targets instead of their corresponding detection boxes in multi-view video sequences. We extract superpixels from detection boxes in all images and associate them over different views and time steps. Association of several superpixels in a detection box results in a segmentation. Moreover, association of several segmentations from different views results in a 3D reconstruction. Finally, association of segmentations or reconstructions over time (i.e., temporal association) results in tracking of segmented targets in video sequences. In other words, we address the problem of segmentation, reconstruction and tracking of multiple targets in multi-view video sequences.

In contrast to previous works, we aim to assign a unique target ID not to each individual detection, but to every superpixel in the entire multi-view video sequence. In common with some other approaches Fragkiadaki and Shi (2011); Hofmann et al. (2013b), the problem is first formulated as a maximum a-posteriori problem and then mapped into a constrained flow graph, which can be efficiently solved by available off-the-shelf binary linear programming solvers. This work is an extension of the work of Hofmann et al. (2013b) which has considered reconstruction in tracking and is inspired by the work of Milan et al. (2015) that addresses both video segmentation and tracking. Our main contributions are 1) combined segmentation, reconstruction and tracking of unknown number of targets in multi-view video sequences; 2) a new constrained flow graph that takes multi-view couplings and low-level superpixel information into account and 3) a novel 'local+global' optimization strategy to simplify the graph and speed up the algorithm while having little impact on the tracking and segmentation performance. Experimental results on standard, publicly available datasets show that the method can outperform many other methods with tracking performance while achieving comparable segmentation performance.

* Corresponding author.
E-mail address: Reza.babaee@tum.de (M. Babaee).

http://dx.doi.org/10.1016/j.cviu.2016.08.006
1077-3142/

2. Related work

Tracking-by-detection is the most successful strategy that has been explored intensively by many researchers Andriluka et al. (2008); Jiang et al. (2007); Leibe et al. (2007); Pirsiavash et al. (2011); Zamir et al. (2012). Here, first a set of detections is obtained by applying object detection algorithms on all images and then is fed into a data association algorithm to track the targets (i.e., finding the identities of targets) in the sequence of frames such that the trajectories of targets are smooth. The main challenge is the data association problem, where the number of possible associations of targets over the time frames increases exponentially with the number of targets. To address this problem, modern approaches cast this problem in different ways such as a graph optimization whose solution can be obtained using Integer Linear Programming Berclaz et al. (2009); 2011), a network flow Pirsiavash et al. (2011); Zhang et al. (2008), continuous or discrete-continuous energy minimization Andriyenko et al. (2012); Milan et al. (2014), and generalized clique graphs Dehghan et al. (2015); Zamir et al. (2012). In order to make the problem tractable, some researchers apply some restrictions, such as reducing the targets' state to the observations in the optimization problem Berclaz et al. (2011); Leal-Taixé et al. (2014); Zhang et al. (2008) or sting measurements Butt and Collins (2013); Leal-Taixé et al. (2014); 2012); Zhang et al. (2008). However, these techniques are only able to track a set of bounding boxes containing the objects. Evidently in many applications finer tracking of the targets is highly desirable.

In order to have finer tracking of objects, video segmentation techniques Bibby and Reid (2010); Brox and Malik (2010); Galasso et al. (2014) are used to assign semantic labels to the pixels in a sequence of frames such that pixels belonging to the same target should preserve their label throughout the entire video sequence. For instance, the authors in Horbert et al. (2011); Mitzel et al. (2010) use video segmentation for pedestrian tracking. Fragkiadaki and Shi (2011) cast the problem of multi-target tracking as clustering of low-level trajectories in order to enhance the tracking results in cluttered situations. Milan et al. Milan et al. (2015) aim to track superpixels over the frame sequence by casting the superpixel tracking as a multi-label optimization problem. They define several types of cost functions in their graphical model (i.e., Conditional Random Field (CRF)). The solution of optimization leads to a joint segmentation and tracking of targets. However, their approach is based on a single view.

The noisy measurements and occlusions greatly influent the performance of single view tracking algorithms since no additional information can be taken advantage of to cope with the difficulties. In order to further improve the tracking quality, researchers proposed many multi-view tracking techniques that leverage different camera views Khan and Shah (2006); Kroeger et al. (2014); Wu et al. (2009). Mikic et al. (1998) apply Kalman Filter for three dimensional object tracking. The world coordinates of the targets are found in a least squares sense. Fleuret et al. (2008) introduce a Probabilistic Occupancy Map (POM) that uses background model to estimate the positions of the targets in each time frame from all views. Then K-Shortest Path (KSP) algorithm is utilized to find the final trajectories. Similarly, Shitrit et al. (2014) employ POM as the people detector. Moreover, an appearance model is introduced to help determine the identities of the targets. The tracking task is then modeled as a multi-commodity network and finally solved using KSP algorithm. Kang et al. (2004) employ Joint Probability Data Association Filter to track multiple moving targets with partial and fully occlusions observed by multiple cameras. A color-based appearance model and a motion model derived from a Kalman Filter define the joint probability of the target. The trajectories are found by maximizing the joint probability model.

Here, we use integer linear programming for joint segmentation, reconstruction and tracking of multiple targets observed by multiple cameras. The proposed approach performs data association among extracted superpixels in each view and also among several views. In our approach, we aim to simultaneously segment, reconstruct, and track targets in a multi-view setup.

3. Approach

As the common methodology introduced in Section 2, we follow the most widely used tracking-by-detection paradigm and first formulate the tracking task as a Maximum a-Posteriori (MAP) problem. As it is very difficult to directly solve a such complicated MAP problem, we follow one solving technique, which maps the MAP formulation to a constrained cost flow graph. In order to get access to the pixel level image information, we split each detection box into superpixels using temporal superpixel algorithm Chang et al. (2013). A superpixel contains a group of pixels that have similar color. This procedure allows us to find the precise shape of the human body rather than a roughly estimated bounding box. These superpixels from the same detection box are again combined together to form a representation of the target, which is called a segmentation. A perfect segmentation should only consist of foreground pixels and precisely describe the shape and pose of the current target. Segmentations from different views can further form a reconstruction, indicating the different observations of the same target from different cameras. At last, the reconstructions (which represent targets in multiple views) and segmentations (which represent targets only in a single view) of the same target are linked along the time, resulting to trajectories. The three-level structure and the whole process are shown in Fig. 1.

The segmentations and the reconstructions are the basic vertices of the flow graph, whose probabilities have a great influence on the segmentation and tracking result. We take advantage of the various aspects of image evidence including geometric cues, appearance and color cues, as well as shape and width/height ratio information to calculate the probabilities of the nodes. The flow costs in the network model are directly related to these probabilities. In this way, we link the MAP formulation of multiple object tracking and the cost flow graph together. A path in the graph represents a feasible trajectory hypothesis, so the paths with overall minimum flow costs ensure the maximum conditional probability, i.e. the solution of the MAP formulation.

In this section, we will introduce our approach in detail. We first specify the MAP formulation of the tracking task under our conditions and then convert it to a cost flow graph model. After proposing the network model, we will define all the probabilities used in the model. Then we will introduce our optimization strategy, including a local optimization which prunes the graph and ensures the tractability of the overall optimization problem, and a global optimization which uses temporal evidence to link graph nodes together, forming trajectories. Finally, we will describe the implementation details such as post-processing procedures.

3.1. MAP formulation

A 2D image detection is defined by a tuple $\mathcal{D}_i = (x_i, s_i, c_i, t_i)$, where x_i is the position, s_i the size of the detection, c_i the camera and t_i the time. A superpixel sp_j is a group of image pixels in one frame that have similar color and each detection $\mathcal{D}_i$ can be split into several superpixels. A segmentation $\mathcal{S}_i = \{sp_j\}$ of a detection $\mathcal{D}_i$ is then a set of at least one superpixel which represent a target (e.g., human body).

$$\mathcal{S}_i \subseteq \{sp | \forall sp_j, sp_k \in \mathcal{S}_i, j \neq k : c_{sp_j} = c_{sp_k} \wedge t_{sp_j} = t_{sp_k}\}. \quad (1)$$