Pattern Recognition Letters 39 (2014) 11–20

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras

Gustavo Führ *, Cláudio Rosito Jung

Universidade Federal do Rio Grande do Sul, Institute of Informatics, Av. Bento Gonçalves, 9500, Setor IV, Porto Alegre 91501-970, Brazil

ARTICLE INFO

Article history:
Available online 11 September 2013

Communicated by S. Sarkar

Keywords:
Patch-based tracking
Weighted vector median filters
Homography
Pedestrian tracking
Pedestrian detection

ABSTRACT

This paper presents a new approach for tracking multiple people in monocular calibrated cameras combining patch matching and pedestrian detection. Initially, background removal and pedestrian detection are used in conjunction with the vertical standing hypothesis to initialize the targets with multiples patches. In the tracking step, each patch related to a given target is matched individually across frames, and their translation vectors are combined robustly with pedestrian detection results in the world coordinate frame using weighted vector median filters. Additionally, the algorithm uses the camera parameters to both estimate the person scale in a straightforward manner and to limit the search region used to track each fragment. Our experimental results indicate that our tracker can deal with occlusions and video sequences with strong appearance variations, presenting results comparable to or better than existing state-of-the-art algorithms.

1. Introduction

Object tracking is an active research topic in the computer vision community. In particular, pedestrian tracking represents an important task in a wide range of applications, such as video analytics, surveillance and analysis of athletic performance in collective sports.

The development of a complete framework for pedestrian tracking involves a variety of challenges. In the initialization step, each new pedestrian that enters the scene must be detected. In general, background removal and/or pedestrian detection algorithms are used in this stage, and they are strongly affected by illumination changes, shadows and varying poses. The tracking phase consists of constantly localizing a pedestrian across time, and it is a complex task due to several reasons: people can move fast and unpredictably, the subject's appearance can change throughout the sequence and the video can present noise and blur, among others. In a common surveillance scenario, the problem can be even harder because of the many occlusions (between a group of people or the person and scene) that can occur. Additionally, these systems often require on-the-fly execution, so non-causal methods that use future information to estimate the current state are not well suited for these applications.

Instead of considering the target as a whole, some approaches split the target into multiple fragments (Adam et al., 2006; Dihl et al., 2011; Führ and Jung, 2012). These approaches have been shown to increase robustness in the presence of partial occlusions, since non-occluded fragments can be matched correctly. This paper extends the fragments-based approach described in Führ and Jung (2012) for pedestrian tracking, by including several new features:

- Multiple people tracking: in this paper, we present an automatic procedure for target initialization based on background removal, pedestrian detection and a vertical standing hypothesis, so that several targets can be detected and tracked simultaneously.
- Use of patch matching and pedestrian detection to improve robustness: the information from a pedestrian detection algorithm is included in the tracker, increasing accuracy and helping to recover the target after occlusions.
- Inclusion of a predicted position based on the history of movement of each target, which provides smoother tracks and helps the tracker during occlusions.

The remainder of this paper is organized as follows. Section 2 reviews some relevant work on object tracking, focusing on pedestrians. The proposed approach is described in Section 3, and the experimental validation is presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

Object tracking is an active research topic in computer vision, and there is a great variety of approaches to tackle the problem.

* Corresponding author. Tel.: +55 (51) 3308 6231; fax: +55 (51) 3308 7308.
E-mail addresses: gfuhr@inf.ufrgs.br (G. Führ), crjung@inf.ufrgs.br (C.R. Jung).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.08.031

This review will focus on pedestrian and/or multi-target tracking, the reader can refer to Yilmaz et al. (2006) for a comprehensive review and taxonomy of generic tracking algorithms. Another useful reference is the survey paper by Enzweiler and Gavrila (2009), which covers the problem of pedestrian detection and tracking using monocular cameras.

A common strategy for pedestrian tracking using static cameras is background removal, which consists of obtaining a mathematical model of the static background, and then comparing each new frame of the video sequence to this model for detecting foreground objects. Detected foreground blobs are candidate pedestrians, and must be validated and tracked across time. For instance, the W4 algorithm (Haritaoglu et al., 1998) uses background segmentation, combined with shape and texture information to perform real time tracking in a monocular gray-scale video. Fleuret et al. (2008) also explore background segmentation, but using multiple calibrated cameras to produce a probabilistic occupancy map of the people in the scene. One drawback of techniques that rely on background removal for the tracking itself is that illumination changes, shadows, noise and clutter may corrupt the extraction of foreground blobs.

Color information is a useful cue to model people appearance and can significantly increase tracking performance. Ramanan et al. (2007) proposed a method to detect and track people using a 2D body model and an appearance model constructed by finding a mean color histogram of different feature candidates extracted in the first few frames of the video. In Fleuret et al. (2008), the intersection between a foreground mask and the frame is used to select the pixels that compose the subject appearance, modeled in a probabilistic framework. A clear drawback of techniques that rely mostly on color is that they cannot be applied to monochromatic video sequences.

In addition to the choice of features, a challenging aspect of pedestrian tracking is to maintain a good localization during and after an occlusion. This is critical for surveillance systems, in which partial and total occlusions can happen very often. In recent years, methods based on data association were proposed for multiple people tracking. Benfold and Reid (2011) use histograms of oriented gradients (HoGs) (Dalal and Triggs, 2005) and Kanade–Lucas–Tomasi (KLT) tracking to detect people and estimate their motion between detections. To obtain the final trajectories, a Markov-Chain Monte-Carlo data association is applied within a temporal window. One drawback of these methods is the latency caused by the use of future observations in the estimation of the current state, i.e. such approaches are not causal.

A different class of approaches is based on tracking-by-detection, which involves the continuous application of a detection algorithm in individual frames and the association of detections across frames. Breitenstein et al. (2011) presented a multi-person online tracking algorithm in an incremental manner: they use class-specific information to detect pedestrians, and also target-specific information to discriminate each pedestrian. Data association across time is performed using a particle filter, using position and velocity to build the state vector. In Kalal et al. (2010, 2012), the authors present an approach that combines detection, learning and tracking. A tracker is used to follow the target in time, while the detector localizes all appearances that have been observed so far and corrects the tracker if necessary and the learning estimates the detector's errors and updates it to avoid these errors in the future. Approaches based on tracking-by-detection present good results when the displacement of the target is large in adjacent frames (e.g. low frame rate video sequences), but may fail when the number of targets increase due to mismatches of independently detected targets. Some of the tracking-by-detection methods perform the temporal association of detections at a current frame using information from future frames (Pirsiavash et al., 2011; Benfold and Reid, 2011). For instance, the approach presented by Pirsiavash et al. (2011) first detects all the pedestrians in the sequence and then uses dynamic programming to associate the detections into trajectories. Usually, these methods present some latency in the estimation or are run completely offline. This is undesired because many surveillance applications require online tracking.

The knowledge of camera information is also useful in pedestrian tracking. Choi and Savarese (2010) proposed a multi-target tracking model to identify the trajectories of multiple objects in 3D based on an initial estimate of the camera parameters. Such 3D trajectories re estimated by measuring their projections onto 2D image plane which represent the observation variables, and then jointly searching the most plausible explanation for both camera and all the existing targets' states in 3D using the projection characterized by the camera model. However, their method assumes that the camera is parallel to the ground plane, being more useful for mobile robots than surveillance using static cameras.

Another way to deal with partial occlusions is to consider the target object as a set of adjacent patches. The rationale behind this idea is that if some patches are occluded and tracked incorrectly, the remaining patches can provide a good estimate of the pose. The FragTrack algorithm (Adam et al., 2006) divides the target region into multiple image fragments at initialization. For each fragment, a vote map is constructed using image histograms. Then, these maps are combined in a robust way so that the influence of outliers is reduced. Dihl et al. (2011) also use the same idea for object tracking, but track each patch independently and combine these tracking results to estimate the location of the target. Führ and Jung (2012) used multiple patches specifically for pedestrian tracking, distributing them along the vertical axis (in world coordinate system – WCS) of each person, and using 3D motion constraints to combine the information of each tracked patch.

The use of multiple fragments has shown good results in generic tracking applications (Adam et al., 2006; Dihl et al., 2011), and also when tailored to pedestrian tracking (Führ and Jung, 2012). This work presents a multiple pedestrian tracker using calibrated cameras that extends the idea presented in Führ and Jung (2012) by including an automatic initialization procedure and the use of a pedestrian detection algorithm to improve tracking results. The proposed approach is presented next.

3. The proposed method

The proposed approach consists of initially detecting the targets (pedestrians), and representing each target as a set of patches. The patches related to each pedestrian are then tracked individually, and their motion patterns are combined in a robust manner in the WCS using a weighted vector median filter (WVMF). A predicted motion vector and a people detector are also included in the tracking framework to improve accuracy and to better handle occlusions. The steps of the proposed method are detailed next.

3.1. Automatic initialization and patch creation

The first step of our approach is to initialize the tracks using a combination of pedestrian detection and background removal in an automatic procedure, which contrasts to our previous work where manual initialization is required (Führ and Jung, 2012). Several algorithms may be used for detecting pedestrians, and in this work we have chosen the method proposed by Dollár et al. (2010), since it presents a good trade-off between accuracy and speed. After the bounding boxes of the pedestrians are found in the image by running this algorithm, the next step is to validate the detection