Signal Processing: *Image Communication* 31 (2015) 100–111

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Spatio-temporal crowd density model in a human detection and tracking framework

Hajer Fradi [a,*], Volker Eiselein [b], Jean-Luc Dugelay [a], Ivo Keller [b], Thomas Sikora [b]

[a] *Multimedia Department, EURECOM, Sophia Antipolis, France*
[b] *Communication Systems Group, Technische Universität Berlin, Germany*

ARTICLE INFO

Article history:
Received 7 March 2014
Received in revised form
25 October 2014
Accepted 21 November 2014
Available online 16 December 2014

Keywords:
Crowd density
Local features
Human detection
Tracking
Crowded scenes

ABSTRACT

Recently significant progress has been made in the field of person detection and tracking. However, crowded scenes remain particularly challenging and can deeply affect the results due to overlapping detections and dynamic occlusions. In this paper, we present a method to enhance human detection and tracking in crowded scenes. It is based on introducing additional information about crowds and integrating it into the state-of-the-art detector. This additional information cue consists of modeling time-varying dynamics of the crowd density using local features as an observation of a probabilistic function. It also involves a feature tracking step which allows excluding feature points attached to the background. This process is favorable for the later density estimation since the influence of features irrelevant to the underlying crowd density is removed. Our proposed approach applies a scene-adaptive dynamic parametrization using this crowd density measure. It also includes a self-adaptive learning of the human aspect ratio and perceived height in order to reduce false positive detections. The resulting improved detections are subsequently used to boost the efficiency of the tracking in a tracking-by-detection framework. Our proposed approach for person detection is evaluated on videos from different datasets, and the results demonstrate the advantages of incorporating crowd density and geometrical constraints into the detection process. Also, its impact on tracking results have been experimentally validated showing good results.

1. Introduction

Automatic detection and tracking of people in video data is a common task in the research area of video analysis and its results lay the foundations of a wide range of applications such as video surveillance, behavior modeling, security applications, and traffic control. Many tracking algorithms use the "Tracking-by-detection" paradigm which estimates the tracks of individual objects based on a previously computed set of object detections. Tracking methods based on these techniques are manifold [1–4], but all of them rely on efficient detectors which have to identify the position of persons in the scene while minimizing false detections (clutter) in areas without people. Techniques based on background subtraction such as [5] are widely applied thanks to their simplicity and effectiveness but are limited to scenes with few and easily perceptible components. Therefore, the application of these methods on videos containing dense crowds is more challenging.

Crowded scenes exhibit some particular characteristics rendering the problem of multi-target tracking more difficult than in scenes with few people: firstly, due to the large number of pedestrians within extremely crowded scenes, the size of a target is usually small in crowds. Secondly, the number of pixels of an object decreases with a higher density due to the occlusions caused by inter-object interactions.

* Corresponding author.
E-mail address: fradi@eurecom.fr (H. Fradi).

http://dx.doi.org/10.1016/j.image.2014.11.006
0923-5965/

Thirdly, constant interactions among individuals in the crowd make it hard to discern them from each other. Finally and as the most difficult problem, full target occlusions that may occur (often for a long time) by other objects in the scene or by other targets. All the aforementioned factors contribute to the loss of observation of the target objects in crowded videos. These challenges are added to the classical difficulties hampering any tracking algorithm such as changes in the appearance of targets related to the camera view field, the discontinuity of trajectories when the target leaves the field of view and re-appears later again, cluttered background, and similar appearance of some objects in the scene.

Because of all these issues, conventional human detection or multi-target tracking paradigms are not scalable to crowds. That is why, some current solutions in crowd analysis field bypass the detection and the tracking of individuals in the scene. Instead, they focus on detecting and tracking local features [6,7], or particles [8,9]. The extracted local features are employed to represent the individuals present in the scene. By this way, tracking of individuals in crowds which is a daunting task is avoided. Likewise, alternative solutions that operate on particles tracking observe that when persons are densely crowded, individual movement is restricted, thus, they consider members of the crowd as granular particles. For instance, in [6], Ihaddadene et al. propose to detect sudden change and abnormal motion variations using motion heat maps and optical flow. The proposed approach is based on tracking points of interest in the regions of interest (masks that correspond to areas of the built motion heat map). Then, the variations of motion are used to detect abnormal events. For this purpose, an entropy measure that characterizes how much the optical flow vectors are organized, or cluttered in the frame is defined in terms of a set of statistical measure. Another study that addressed the problem of abnormal crowd event detection is the social force model proposed by Mehran et al. [8]. This method is based on putting a grid of particles over the image frame and moving them with flow field computed from the optical flow. Then, the interaction forces are computed on moving particles to model the ongoing crowd behaviors. In the same context of crowd behavior analysis, other methods [10,11] studied the dynamic evolution of the crowd using biological models.

Most of the proposed works to tackle multi-target tracking in crowded scenes use motion pattern information as priors to tracking. Some of these methods are applied in unstructured crowd scenes [12], while most of them focus on structured scenes [13–15], where objects do not move randomly, and exhibit clear motion patterns. In [12], a tracking approach in unstructured environments, where the crowd motion appears to be random in different directions over time is presented. Each location in the scene can represent motion in different directions using a topical model. In [13], a motion structure tracker is proposed to solve the problem of tracking in very crowded scenes. In particular, tracking and detection are performed jointly and motion pattern information is integrated in both steps to enforce scene structure constraints. In [14], a probabilistic method exploiting the inherent spatially and temporally varying structured pattern of crowd motion is employed to track individuals in extremely crowded scenes. The spatial and temporal variations of the crowd motion are captured by training a collection of hidden Markov models on the motion patterns within the scene. Using these models, pedestrian movement at each space–time location in a video can be predicted. Also motion patterns are studied in [15], where floor fields are proposed to determine the probability of moving from one location to another. The idea is to learn global motion patterns and participants of the crowd are then assumed to move in a similar pattern. Finally, in [16] a spatiotemporal viscous fluid field is proposed to recognize large-scale crowd event. In particular, a spatiotemporal variation matrix is proposed to exploit motion property of a crowd. Also, a spatiotemporal force field is employed to exploit the interaction force between the pedestrians. Then, the spatiotemporal viscous fluid field is modeled by latent Dirichlet allocation to recognize crowd behavior.

Although these solutions have shown promising results, they impose constraints on the crowd motion. In particular, targets are often assumed to behave in a similar manner, in such a way that all of them follow a same motion pattern, consequently, trajectories not following common patterns are penalized. Certainly, this constraint works well in extremely crowded scenes, such as in some religious events or demonstrations, where the movement of individuals within the crowd is restricted by others and by the scene structure as well. Thus, a single object can be tracked by the crowd motion because it is difficult, if not impossible, to move against the main trend. However, the aforementioned methods are not applicable in cases where individuals can move in different directions. Furthermore, some of these methods include other additional constraints. For example, in [12], Rodriguez et al. use a limited descriptive representation of target motion by quantizing the optical flow vectors into 10 possible directions. Such a coarse quantization limits tracking to only few directions. Also, the *floor fields* [15] used by Ali et al. impose how a pedestrian should move based on scene-wide constraints, which results in only one single direction at each spatial position in the video.

In addition to these solutions based on exploiting global level information about motion patterns to impose constraints on tracking algorithms, similar ideas have been proposed using crowd density measures. In [17], Hou et al. use the estimated number of persons in the detection step, which is formulated as a clustering problem with prior knowledge about the number of clusters. This attempt to improve person detection in crowded scenes includes some weaknesses. At least two problems might incur: firstly, the idea of detection by clustering features can only be effective in low crowded scenes. It is not applicable in very crowded cases because of the spatial overlaps that make delineating individuals a difficult task. Secondly, using the number of people as a crowd measure has the limitation of giving only global information about the entire image and discarding local information about the crowd.

We therefore resort to another crowd density measure, in which local information at pixel level substitutes a global number of people per frame. This alternative solution based on computing crowd density maps is indeed more appropriate as it enables both the detection and the location of potentially crowded areas. To the best of our knowledge, only one work [18] has investigated this idea. In the referred work, a system which introduces crowd density information