Neurocomputing 177 (2016) 543–563

Contents lists available at ScienceDirect

Neurocomputing

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

Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows

Sultan D. Khan [a], Stefania Bandini [a,c], Saleh Basalamah [b], Giuseppe Vizzari [a,*]

[a] *Complex Systems and Artificial Intelligence Research Centre, Università degli Studi di Milano–Bicocca, Milano, Italy*
[b] *Department of Computer Engineering, Umm Al Qura University, Makkah, Saudi Arabia*
[c] *RCAST, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan*

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form
19 September 2015
Accepted 17 November 2015
Communicated by Marco Cristani
Available online 3 December 2015

Keywords:
Crowd analysis

ABSTRACT

Pedestrians, in videos taken from fixed cameras, tend to appear and disappear at precise locations such as doors, gateways or edges of the scene: we refer to locations where pedestrians appear as sources (potential origins) and the locations where they disappear as sinks (potential destinations). The detection of these points and the characterization of the flows connecting them represent a typical preliminary step in most pedestrian studies and it can be supported by computer vision approaches. In this paper we propose an algorithm in which a scene is overlaid by a grid of particles initializing a dynamical system defined by optical flow, a high level global motion information. Time integration of the dynamical system produces short particle trajectories (tracklets), representing dense but short motion patterns in segments of the scene; tracklets are then extended into longer tracks that are grouped using an unsupervised clustering algorithm, where the similarity is measured by the Longest Common Subsequence. The analysis of these clusters supports the identification of sources and sinks related to a single video segment. Local segment information is finally combined to achieve a global set of traces identifying sources and sinks, and characterizing the flow of pedestrians connecting them. The paper presents the defined technique and it discusses its application in a real-world scenario.

1. Introduction

Crowded scenes are composed of a large number of people, exhibiting different behaviors in a relatively constrained space. The vagueness of this definition is strictly related to the difficulties in defining what a crowd of pedestrian is; we will not try here to be more specific or precise, but rather highlight the growing need to consider the presence and behaviors of pedestrians in the environment by designers, planners and decision makers (see, e.g., a recent report commissioned by the U.K. Cabinet Office on this subject [1]). In particular, public safety in crowded situations (e.g. concerts, religious or political gatherings) has become an important research area in the last years, with relevant contributions from physics, psychology, computer science and, of course, civil engineering. Acquiring data for this kind of study is obviously absolutely crucial for sake of understanding the implied phenomena and evaluating developed solutions for analysis, decision support, prediction. In video surveillance, scene modeling and understanding is also an important research area. Important tasks of scene modeling and understanding are (i) extracting motion information (e.g. trajectories), (ii) identification of entry and exit points of trajectories in the analyzed scene, (iii) characterization of the interaction of trajectories (highlighting, for instance, crossings or potential conflicts).

Pedestrians in videos taken from fixed cameras tend to appear and disappear at relatively precise and recurring locations, such as doors, gateways or particular portions of the edges of the scene. Moreover, pedestrian behavior in a given scene might imply waiting at a certain location then moving whenever certain conditions are met or given events happen. We refer to locations where pedestrians appear or start moving as *sources* (potential origins of a trajectory) and the locations where they disappear or stop moving as *sinks* (potential destinations). Traditionally, crowd analysis is performed by the analyst who manually identifies and detects different relevant activities in the scene. A portion of video is given to each analyst together with a list of events (behaviors) and objects to look for. The analyst informs the concerned authorities if any of the given events or objects are detected. Such kind of manual analysis of video is labor intensive, time consuming and prone to errors due to weak perceptive capabilities of humans, but also to the repetitiveness of the activity.

[*] Corresponding author.
E-mail addresses: sultan.khan@disco.unimib.it (S.D. Khan), bandini@disco.unimib.it (S. Bandini), smbasalamah@uqu.edu.sa (S. Basalamah), vizzari@disco.unimib.it (G. Vizzari).

http://dx.doi.org/10.1016/j.neucom.2015.11.049
0925-2312/© 2015 Elsevier B.V. All rights reserved.

In this paper, we propose an approach for crowd behavior analysis (and, to a certain extent, understanding in the acceptation of the term adopted by Junior et al. [2]) adopting two novel algorithms, the first able to generate long, dense, reliable and accurate pedestrian trajectories and the second clustering them to generate long term reliable and abstract information describing flows in the whole video. The final results provide directly information characterizing flows but it also represents a starting point for further high-level analyses of crowd behavior. The approach starts by dividing the input video into multiple *segments* of equal length and, considering that the frame rate of the video is constant, duration. The initial frame of each segment is overlaid by a grid of particles initializing a dynamical system defined by optical flow, as discussed by Solmaz et al. [3]. Time integration of the dynamical system over a segment of the video provides particle trajectories (tracklets) that represent motion patterns in the scene for a certain time interval associated to the analyzed segment. We detect sources, sinks and main flows in the segment (for sake of brevity sometimes we will refer to this information as segment local *track*) by analyzing motion patterns followed by clusters of tracklets, obtained using an unsupervised hierarchical clustering algorithm, where the similarity is measured by the Longest Common Sub-sequence (LCS) metric. To achieve final global tracks, covering all the video, we cluster the achieved local tracks through the same hierarchical clustering algorithm. Our main contributions are: (1) Generating dense and long trajectories, (2) identifying sources and sinks, (3) understanding behavior of the crowd in the scene by considering full length video, (4) achieve the above results without requiring object detection, tracking, nor training, targeting employment in naturalistic conditions. The paper breaks down as follows: the following section presents the current state of the art in the identification and characterization of pedestrian flows in crowded scenes, while Section 3 presents the overall proposed approach. Section 3.1 focuses on the algorithm to extract long, dense, accurate and reliable trajectories and Section 3.4 describes in details the clustering algorithm applied to generate local and global tracks. Section 4 describes the achieved experimental results, also by comparing the proposed approach with the most relevant existing alternatives. Conclusions and future developments end the paper.

2. Related works

With the advancement in computer vision technology, researchers developed tracking methods that in certain conditions can automatically detect, track and identify specific activities in the scene. Zhao and Nevatia [4] developed a tracking algorithm by modeling human shape and appearance as articulated ellipsoids and color histogram respectively for crowded scenes. Khan et al. [5] use Markov chain Monte Carlo based particle filter to handle interaction between multiple targets in crowded scene. Hue et al. [6] detect interest points in each frame by tracking pedestrians, and this activity is performed by finding correspondence among points between frames. Brostow and Cipolla [7] developed an unsupervised bayesian clustering method to detect individuals in crowd: for each frame, detection of individuals is performed ignoring the relationship between frames. Sugimura et al. [8] detect individual objects on the basis of assumption that objects move in different directions. Zamir et al. [9] develop a tracking system by solving data association problem by utilizing Generalized Minimum Clique Graph (GMCP) in order to detect an individual in different frames of a video. Intuitively, detection and tracking of individuals rely on the performance of detection and tracking algorithms. However, in crowded scenes, where the number of objects is often in the order of hundreds, these tasks usually fail due to (i) the variable and potentially low number of pixels per object and (ii) frequent and severe occlusions related to the constant interaction among the objects (pedestrians) in the scene. These challenging characteristics of the analyzed videos can be at least partly avoided in laboratory situations: for instance, in [10] the authors successfully gather pedestrians' trajectories and gather useful data about their behavior but they employ a manual or automatic but facilitated form of identification. Moreover, as we will discuss in the experimental evaluation of the presented approach, the adopted tracking algorithm (Lucas-Kanade tracker - KLT [11]) does not provide sufficiently accurate results in naturalistic conditions.

Intuitively, detecting sources and sinks (as introduced above) implies detection and tracking of objects, potentially followed by an analysis of the trajectories: this kind of approach was adopted by Stauer [12], which analyses low density situations and essentially relies on the performance of the tracking algorithm, which is low in crowded situations. Research in this area has therefore instead assumed that raw data about pedestrian paths should be considered as noisy or unreliable: [13], for instance, employ a so-called *weak tracking* system and they aggregate raw *tracklets* through a mean-shift clustering technique allowing them to identify entry and exit zones in the scene. More recently, in order to overcome the limitation of traditional tracking methods, research has focused on gathering global motion information at higher level, often based on optical flow analysis.

Trajectories capture the local motion information of the video. Long and dense trajectories (that is, trajectories representing a large number of paths followed by different pedestrians, reaching a significant length) provide good coverage of foreground motion as well as of the surrounding context. There are two types of representations for characterizing motion information from the video: space-time local features (like corner points, SIFT features, etc.) and dense optical flow [14]. In the first type, features are detected in one frame which are then tracked through rest of the frames of a video, whereas the second type is based on dense optical flow, where a flow vector is estimated for every pixel. Since dense optical flow estimates a change for every pixel it provides a better representation of motion in video. A large number of approaches for extracting feature trajectories from video exist:

- the work described in [15] extracts feature trajectories by tracking Harris3D interest points; Laptev [16] used KLT for extracting trajectories represented as a sequence of log-polar quantized velocities which later on used for action classification;
- a different approach [17] also used KLT for extracting trajectories, that are then clustered and affine transformation matrix representing trajectories is computed for each cluster;
- other researchers extract trajectories by matching SIFT descriptors between two consecutive frames [18];
- the work described in [19] combine both KLT tracker and SIFT descriptor matching to extract long-duration trajectories, and random points are sampled for tracking within the region of existing trajectories in order to assure dense coverage;
- another approach [20] extracts feature point trajectories in the regions of interest; in this work, authors compute histogram of gradient (HOG) and histogram of optical flow (HOF) descriptors along the trajectories.
- KLT method is also used in [21] for extracting sparse trajectories: the authors propose Random Topic Model (RTM) for learning semantic regions from the motions of pedestrians in crowds. A variant of this approach [22] employs KLT trajectories and proposes Mixture model of Dynamic pedestrian Agents (MDA) that analyse the collective behavior of pedestrian in crowds after learning from the real data.