Digital Signal Processing 57 (2016) 106–127

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A numerical approach for tracking unknown number of individual targets in videos

Md. Haidar Sharif

International University of Sarajevo, Bosnia and Herzegovina

ARTICLE INFO

Article history:
Available online 14 July 2016

Keywords:
Big O notation
Brute-force search
Hungarian method
Kalman filter
Tracking
Videos

ABSTRACT

Suppose that we wish to get a comprehensive match of a target in the next frame. Where would we search the target in the next frame? Brute-force search has an asymptotic runtime of $O(n!)$ with problem size n. Yet we can search the target only from a number of automatically generated specific regions, named candidate regions, in the next frame. But how can we get those regions? Deeming the silhouettes of movers, this paper denotes a detailed deliberation of how to generate those candidate regions automatically and then how to track unknown number of individual targets with them. Phase-correlation technique aids to find the key befitting matches of the targets using them. Hungarian method in combination with a state estimation process called Kalman filter finds the best correspondence of the targets among those matches, allowing us to construct full trajectories of unknown number of individual targets in 3D space irresistibly swift as compared to brute-force search since the relative runtime reduced from $O(n!)$ to $O(n^3)$. Favorable outcomes, upon conducting experiments on videos from three different datasets, show the robustness and effectiveness of our approach.

1. Introduction

Target tracking, which aims at detecting the position of a moving object from video sequences, is a challenging research topic in computer vision. Obstacles in tracking targets can grow due to quick target motion, changing appearance patterns of target and scene, nonrigid target structures, dynamic illumination, inter-target and target-to-scene occlusions, and multi-target confusion. As selection of features affects the tracking results, it is eventful to select right features. Feature selection is closely related to the target representation. A target shape can be represented by a primitive geometric shape including rectangle, ellipse, circle, square, triangle, and point [1]. The efforts of tracking targets or objects in videos as efficient as possible are not new [1–25]. A vast majority of the existing algorithms primarily differ in the way they use image features and model motion, appearance and shape of the target. For instance, in silhouette-based tracking, silhouettes are tracked by either shape matching [4] or contour evolution [2]. In kernel-based tracking, target can be tracked by computing the motion of the kernel in consecutive frames [7,8]. Kernel can be an elliptical shape with an associated histogram [3] followed on a mean-shift procedure to locate the target or a rectangular template with an associated covariance matrix [6] come after a brute-force search for locating the target. Both silhouette-based and kernel-based tracking approaches may arouse more attention. But widely due to algorithmic assumptions, variety in appearance of target with various view angles, and divergent degrees of imperfect occlusion many of them would not be workable to track targets individually in videos. Thus the recent line of trends are differing from the previous trends by developing more efficient approaches capable of handling crowded scenes by chiefly fixing on deriving a robust appearance model of each individual. For instance, Ali et al. [7], Rodriguez et al. [8], and Kratz et al. [15] handled crowded scenes by largely focusing on deriving robust appearance models of each single. Ali et al. [7] computed dynamic floor field using frames after current tracking position, while Kratz et al. [15] used previously observed frames in video to predict the motion of target. Rodriguez et al. [8] used a topical model to show motion in different directions at each spatial location. Their approach imposed a fixed number of motion directions at each spatial location, but disregarded the temporal relationship between sequentially occurring local motions. In contrast, Kratz et al. [15] expressly encoded this relationship with a hidden Markov model at each spatial location in video. In addition, Rodriguez et al. [8] quantized the optical flow vectors into 10 possible directions, while Kratz et al. [15] used a full distribution of optical flow for a more robust and descriptive depiction of

E-mail address: haidar@ius.edu.ba.

http://dx.doi.org/10.1016/j.dsp.2016.07.004
1051-2004/

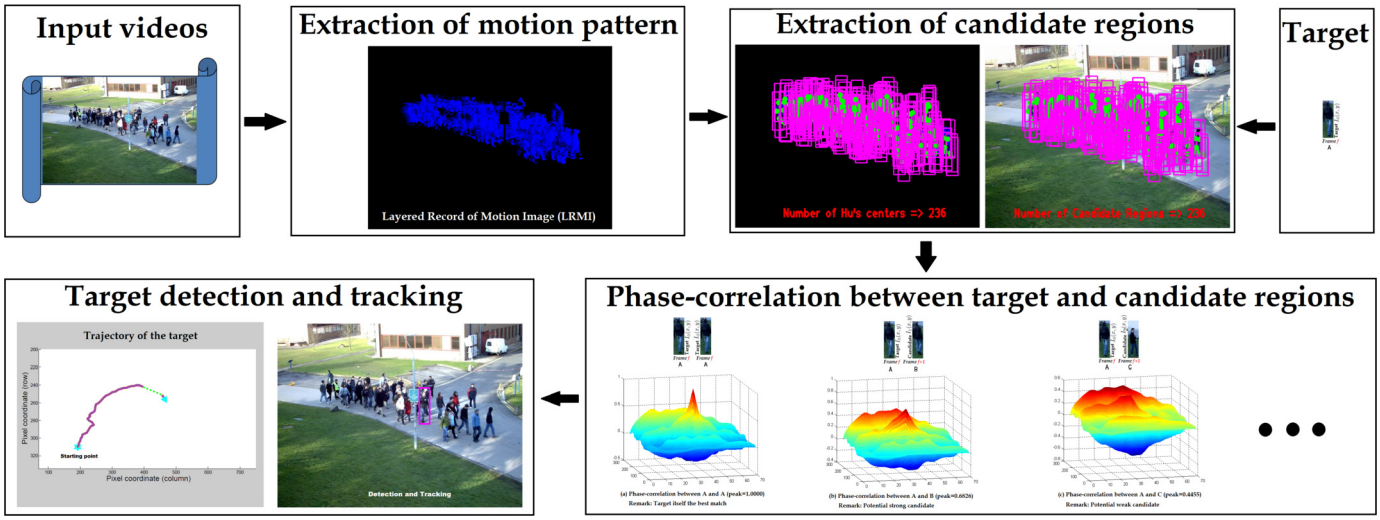

Fig. 1. Summary of our approach.

the motion of tracked targets. Such a coarse quantization limited tracking in crowded scenes to only a few directions [15].

In this paper, we have tracked unknown number of individual targets in videos by leveraging a spatiotemporal motion model of the movers. We have addressed an innovative idea of how to extract the candidate regions, irrespective of movers' number in the scenes, from the silhouetted structures of movers' pixels. Silhouettes of movers obtained on capturing the local spatiotemporal motion patterns of movers, we generate candidate regions in the candidate frames in a reflex manner. Candidate frame means a frame where a target template would be matched with one of its available movers. Our approach would fall in the category of silhouette-based template matching and tracking in the following candidate frames in some sense. Suppose that we wish to get a comprehensive match of a target in the candidate frame. Immediately, a common question will be come to pass: Where will we search the target in the candidate frame if the candidate regions are not known to us? One possible answer would be the integral images [26,5], which will allow for very fast feature evaluation. Even though each feature can be computed very efficiently, computing the complete set is prohibitively expensive. For example, if the base resolution of a detector is 24×24 pixels, the exhaustive set of rectangle features is over 180 000, which is far larger than the number of pixels [26]. With effective classifier integral images would provide a minimum solution of that problem. The search of a target throughout the whole frame starting from either upper/lower leftmost or rightmost corners is a time consuming task for a single target let multiple targets alone. With problem size n, the brute-force search holds asymptotic runtime of $O(n!)$, which causes the unwanted phenomenon of combinatorial explosion for many real-world problems. To solve this consternation, we have proposed candidate regions in the candidate frame. We can search the target solely some automatically generated specific regions, called candidate regions, in the candidate frame. Based on the shape and size of the track of interest, candidate regions can be represented by rectangles, ellipses, circles, squares, and triangles. But how can we have a firm footing of candidate regions in the candidate frame automatically? This paper makes ready to a detailed formal exposition of how to generate those candidate regions followed by the techniques of matching and tracking unknown number of individual targets in the following candidate frames. Fig. 1 briefly gives the gist of our proposed approach.

Our central concept is the Layered Record of Motion Image (LRMI), which is based on the extraction of local spatiotemporal motion patterns using numerical values $\omega = \{0, 1, 2, 3, \ldots, 255\}$ assigned to the movers' silhouettes. Segmentation of LRMI by clustering gives the basic knowledge of candidate regions automatically. Centroid or Hu's center of each segment is estimated by Hu's moments [27]. Each Hu's center is used to present with a fixed size primitive geometric shape on the camera view frame to show each candidate region. In some crowded scenes target shape is very big or whole body may not be visible at all over the duration of the video. On such cases, instead of the complete body of target, a contiguous most visible part of the target can be the track of interest. Using candidate regions phase-correlation technique helps to find the main suitable matches (i.e., strong candidates) of the target on deleting weak candidate regions. And then Hungarian method [28,29] in combination with a state estimation technique called Kalman filter [30] find the best correspondence of the target from strong candidates, allowing us to construct target's full trajectory in 3D space (x-coordinate, y-coordinate, and frames or time). Theoretical runtime of our tracking algorithm is $O(n^3)$. So it tracks targets in polynomial time of degree 3, whereas brute-force search counts on $O(n!)$. The video sequences of PETS-2012 benchmark dataset [31], Web dataset (collected from Shutterstock [32] containing various videos of fishes, birds, ants, bacteria, and red blood cells), and UCF crowd dataset [33] have been esteemed as performance evaluation.[1] Credible results got from the experiments on these datasets show that our method gains high-quality results to track individual targets in crowd videos in terms of both robustness and effectiveness. Reported results using same videos made clear and visible that our method is likely a bit superior to alternative methods (e.g., Ali et al. [7], Rodriguez et al. [8], and Kratz et al. [15]).

Compendiously, the cardinal contributions of this paper are two folded: (i) How to generate candidate regions automatically from video frames; (ii) Unknown number of individual targets tracking technique based on Hungarian method and Kalman filter using candidate regions has been proposed. The framework is easy to implement. One of its key boons is that we do not need to search target regions everywhere in the candidate frame except candidate regions. Accordingly, in search process it leaves out some well-established concepts (e.g., integral images [5] and brute-force search) and becomes overpoweringly rapid as compared to brute-

[1] Readers are invited to watch the supplementary video files for system's tracking performance.