Neurocomputing 100 (2013) 86–97

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Human behavior analysis in video surveillance: A Social Signal Processing perspective

Marco Cristani [a,b], R. Raghavendra [a,*], Alessio Del Bue [a], Vittorio Murino [a,b]

[a] Istituto Italiano di Tecnologia (IIT), Genova, Italy
[b] Dipartimento di Informatica, University of Verona, Italy

ARTICLE INFO

Available online 11 May 2012

Keywords:
Video surveillance
Social Signal Processing
Activity recognition
Behavior analysis
Human computing

ABSTRACT

The analysis of human activities is one of the most intriguing and important open issues for the automated video surveillance community. Since few years ago, it has been handled following a mere Computer Vision and Pattern Recognition perspective, where an activity corresponded to a temporal sequence of explicit actions (run, stop, sit, walk, etc.). Even under this simplistic assumption, the issue is hard, due to the strong diversity of the people appearance, the number of individuals considered (we may monitor single individuals, groups, crowd), the variability of the environmental conditions (indoor/outdoor, different weather conditions), and the kinds of sensors employed. More recently, the automated surveillance of human activities has been faced considering a new perspective, that brings in notions and principles from the social, affective, and psychological literature, and that is called Social Signal Processing (SSP). SSP employs primarily nonverbal cues, most of them are outside of conscious awareness, like face expressions and gazing, body posture and gestures, vocal characteristics, relative distances in the space and the like. This paper is the first review analyzing this new trend, proposing a structured snapshot of the state of the art and envisaging novel challenges in the surveillance domain where the cross-pollination of Computer Science technologies and Sociology theories may offer valid investigation strategies.

1. Introduction

Since the 1990s, human activity analysis has been one of the most important topics in computer vision, becoming an integral part of many video surveillance systems, but also representing a key application in several other everyday scenarios like workplaces, hospitals, and many others. Analyzing activities involved to date the recognition of motion patterns, and the production of high-level descriptions of actions and interactions among entities of interest. Many surveys on activity analysis have been proposed in the literature: the first example is [1], where techniques for the tracking and the recognition of human motion are reviewed; in [2], methods for the motion of body parts, the tracking of human motion using different camera settings and the recognition of activities are reported. In [3], hand and body tracking strategies are discussed, together with techniques for human activity recognition based on 2D and 3D models. A comprehensive review on vision-based human motion analysis spanning the period 2000–2006 is presented in [4]. In [5], statistical models like Dynamic Bayesian Networks are addressed as one of the most suitable tools for activity recognition. An essay on the different components of a typical video surveillance system, with emphasis on the activity analysis, is reported in [6]. The definition of activity as a complex and coordinated organization of simple actions is exploited in [7]. In the same year, a survey on video surveillance systems has been proposed in [8], also discussing about the different public databases available to validate the algorithms. In the very recent review on activity recognition approaches [9], the different strategies are organized as hierarchical and nonhierarchical, and the last ones are further divided into space–time and sequential methods.

All the above-mentioned surveys addressed the modeling of the human activities mainly stressing the technological computer vision aspects. In particular, all of them focus on detecting and recognizing explicit actions, in the sense of gestures performed voluntarily by humans, like running, walking, stopping, seating, etc.

Recently, the study on human activities has been revitalized by addressing the so-called *social signals* [10], which are nonverbal cues inspired by the social, affective, and psychological literature [11]. This allows a more principled encoding of how humans act and react to other people and environmental conditions. Social Signal Processing (SSP), also named Social Signaling, represents the scientific field aimed at a systematic, algorithmic and computational analysis of social signals, that is deeply rooted in

* Corresponding author.
E-mail address: raghu07.mys@gmail.com (R. Raghavendra).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2011.12.038

anthropology and social psychology [12]. More properly, SSP goes beyond the mere human activity modeling, aiming at coding and decoding the human *behavior*. In other words, it focuses to unveil the underlying hidden states that *drive* one to act in a determined way, with particular actions. This challenge is motivated by decades of investigation in human sciences (psychology, anthropology, sociology, etc.) that showed how humans use nonverbal behavioral cues like facial expressions, vocalizations (laughter, fillers, back-channel, etc.), gestures or postures to convey, *often outside conscious awareness*, their attitude towards other people and social environments, as well as emotions [13]. The understanding of these cues is thus paramount in order to understand the social meaning of the activities.

As we will see later, only a minority of works adopted the SSP perspective in a video surveillance setting, but recently (i.e., since 5 years) this trend has rapidly grown. Actually, in surveillance, the main goal is to detect threatening actions as soon as possible: therefore, the possibility of doing this by observing the human behavior as a phenomenon subjected to rigorous principles that produces predictable patterns of activities, turns out to be incredibly important.

The aim of this paper is to review the early years of the social signaling oriented approaches for human behavior analysis in a surveillance context, individuating what are the contact points between surveillance and social signalling, how social signalling may improve the human behavior analysis, envisaging and delineating future perspectives.

The rest of the paper is organized as follows. Section 2 illustrates the processing scheme of a typical video surveillance system. The aim of the section is that of contextualizing which modules of a video surveillance strategy may benefit from the intervention of Social Signaling findings. Section 3 is a short overview of the recent advances in the activity analysis, aimed at defining what is achieved with pure Computer Vision and Pattern Recognition methods. Section 4 is the core of the paper, reviewing the most significant contributions that represent the intersection between video surveillance and SSP. Section 5 addresses the analysis of crowd behavior, that recently has become a well-defined trend in surveillance, discussing the importance of embedding social signals in such studies. Finally, Section 6 draws the conclusions and presents the envisaged future perspectives.

2. A basic video surveillance system overview

A typical surveillance system scheme is composed of two parts: a low-level and a high-level part (see Fig. 1). Each part is composed of different stages, explained in the following.

2.1. The low-level stages

The low-level stages are the background subtraction/object segmentation and the object detection. Such stages preprocess the raw images in order to discover areas of interest.

Background subtraction/object segmentation: Background (BG) subtraction is a fundamental low-level operation that applies on raw videos captured by CCTVs [14]. It aims at learning the expected chromatic aspect of the scene and how it evolves in time, highlighting moving objects (foreground, FG), ideally under a 24/7 policy. Object segmentation follows the background subtraction and aims at individuating connected regions, pruning away small FG objects, filling holes of large regions, adopting temporal continuity to obtain consistent, smooth regions across time [15].

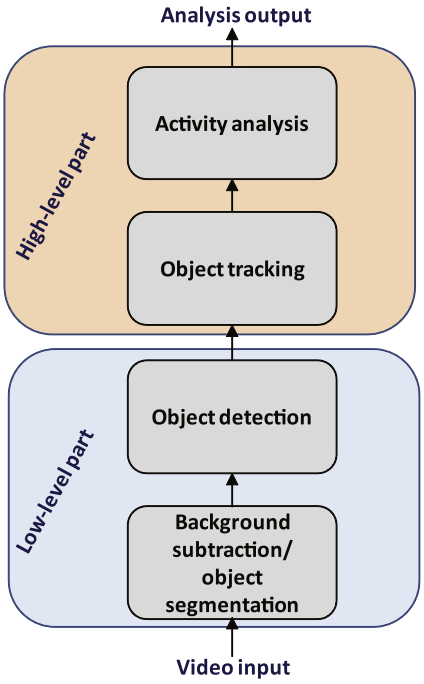

Fig. 1. Typical video surveillance automated system.

Object detection: This stage serves to highlight particular classes of targets (humans, vehicles, baggages) in the images. It may be applied on the output of the background subtraction/object segmentation step, or in a dense way over the entire image [16].

These two stages cannot benefit of an intervention of SSP principles, since the processing here is focused on entities, the pixels, carrying very low semantics.

2.2. The high-level stages

The high level stages are the object tracking and the activity analysis.

Object tracking: Tracking is undoubtedly the paramount aspect of any video-surveillance approach, and is very important for the human behavior analysis. For a comprehensive review on tracking for surveillance (out of the scope of this contribution), read [17]. Tracking aims at computing the trajectory of each distinct object of interest in the scene, associating an ID label and keeping it across occlusions and multiple cameras. A general tracker can be characterized by three main phases: (1) the initialization phase localizes the target that needs to be tracked. It usually relies on heuristic mechanisms combined with some object detector. (2) The dynamic phase predicts where target is more likely to move, and it is based usually on a first- or second-order autoregressive model. (3) The observation phase finds the region of the image that is more similar to the target, assuming as prior the hypothesis given in the dynamical phase.

Tracking, and especially the dynamic module, may benefit from Social Signal Processing methods. Such module simply does not take into account that people, whenever free to move in a large environment (e.g., the hall of a hotel, a square, a waiting room, etc.), respect patterns and trajectories largely dominated by social mechanisms [18]. Therefore, the design of a socially driven dynamic model for tracking may be the key ingredient to overcome the current limitations of the current algorithms, as already shown in some recent approaches exploiting the Social Force Model [19,20] (see later for further details) . When the scenario is too crowded, so that tracking approaches become ineffective, motion flow estimation techniques are usually preferred [21].