Expert Systems with Applications 42 (2015) 6935–6956

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Review

Automatic visual detection of human behavior: A review from 2000 to 2014

Palwasha Afsar *, Paulo Cortez, Henrique Santos

ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal

ARTICLE INFO

Article history:
Available online 22 May 2015

Keywords:
Data mining
Human behavior
Literature review
Video analysis
Video databases

ABSTRACT

Due to advances in information technology (e.g., digital video cameras, ubiquitous sensors), the automatic detection of human behaviors from video is a very recent research topic. In this paper, we perform a systematic and recent literature review on this topic, from 2000 to 2014, covering a selection of 193 papers that were searched from six major scientific publishers. The selected papers were classified into three main subjects: detection techniques, datasets and applications. The detection techniques were divided into four categories (initialization, tracking, pose estimation and recognition). The list of datasets includes eight examples (e.g., Hollywood action). Finally, several application areas were identified, including human detection, abnormal activity detection, action recognition, player modeling and pedestrian detection. Our analysis provides a road map to guide future research for designing automatic visual human behavior detection systems.

1. Introduction

Human detection and their corresponding behaviors have been studied under distinct perspectives in a wide variety of disciplines, such as psychology, biomechanics and computer graphics (Ko, 2008). According to Moeslund, Hilton, and Krüger (2006) and Poppe (2010) an action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed. The main research question underneath this work can be characterized as follows: given a succession of pictures with one or more persons performing an action, can a framework be outlined to distinguish: *who* is performing the action and *what* action was performed? In this paper, we focus on computational frameworks for the automatic detection of human behavior, which is a very recent and relevant research area due to its potential impact in a wide variety of human activities, such as gaming, visual surveillance and detection of elderly accidents.

The particular interest in this area has dramatically increased with the advances of information technology. In particular, depth-imaging (3D) has substantially improved in the last few years, finally reaching an affordable consumer price (e.g., Kinect for Xbox 360). This is a multidisciplinary area, involving several fields such as artificial intelligence, data mining, psychology, biomechanics, pattern recognition and image processing. In effect, the combination of all these fields is necessary to tackle challenging tasks, such as real-time segmentation of changing scenes in natural environments, and non-rigid movement and self-occlusion.

While this is considered a trendy subject, whose interest will increase in the next few years, within our knowledge there are few recent review articles that cover well this area. Some surveys only target a portion of visual human behavior detection possibilities, such as: Moeslund et al. (2006), which analyzed human motion capture from 2000 to 2006; Yampolskiy and Govindaraju (2008), which focused on behavioral biometrics from 1989 to 2007; and Ko (2008), which addressed video surveillance from 1985 to 2008.

Other surveys closely related to action and activity recognition has been done by Turaga, Chellappa, Subrahmanian, and Udrea (2008) and Poppe (2010). The review of Turaga et al. (2008) focus on high-level recognition of actions and activities (e.g., bending, walking, shaking hands) from 1988 to 2008, while the study of Poppe (2010) surveys low-level image representations for action recognition (e.g., optical flow, spatial–temporal volume) from 1992 to 2009. In this paper, we perform a systematic and more recent review (from 2000 to 2014) that includes both low-level and high-level methods for human behavior recognition. Moreover, we also compare the distinct methods and highlight

* Corresponding author.
E-mail addresses: palo_afsar77@yahoo.com (P. Afsar), pcortez@dsi.uminho.pt (P. Cortez), hsantos@dsi.uminho.pt (H. Santos).

http://dx.doi.org/10.1016/j.eswa.2015.05.023
0957-4174/

their advantages and limitations. Finally, we also present several updated examples of applications and the most commonly used datasets (with best performances so far achieved) in this topic.

A comprehensive search was made by analyzing recent papers, from 2000 to 2014, from high quality journals and conferences related with the addressed topic (e.g., Expert Systems with Applications, Computer Vision). The search was executed using six general scientific databases: ACM Digital Library (dl.acm.org), Elsevier (www.elsevier.com), IEEE Xplore Digital Library (ieeexplore.ieee.org), Springer Link (link.springer.com), MIT Press (mitpress.mit.edu) and Wiley Online Library (onlinelibrary.wiley.com). The list of keywords used in the search included combinations of the keywords "Human Behavior" or "Human Detection" with "Video" or "Data". The search was then filtered manually and reduced to 193 papers related to this review. The selected papers were classified into three main subjects: detection techniques, datasets and applications. Techniques that were related to detection were further divided into four categories, namely initialization, tracking, pose estimation and recognition. Moreover, a total of eight datasets were identified (e.g., Hollywood action). Finally, the applications were grouped into six main areas: human detection, abnormal activity detection, action recognition, player modeling and robotics, pedestrian detection and in-home scenarios, and person tracking. To summarize the review in terms of topics and their publishing year, Table 1 presents the evolution of the surveyed papers keywords that appear with five or more occurrences from 2000 to 2014.

This paper is organized as follows. Firstly, Section 2 introduces the visual detection of human behavior computational techniques. Then, Section 3 presents the main datasets used within the surveyed domain. Next, Section 4 lists examples of relevant human behavior detection applications. Finally, Section 5 concludes the review by performing a global analysis to the presented review and presenting future research implications.

2. Techniques used for human behavior detection from video

Following the works of Moeslund and Granum (2001) and Moeslund et al. (2006), the visual human behavior detection techniques were categorized into four groups:

- **Initialization** – in order for the system to process data, it needs to be initialized; e.g., a proper model of the system should be built;
- **Tracking** – the process of segmenting the subjects from the background and finding correspondences between segments in consecutive frames;
- **Pose** – the estimation of pose is carried out in corresponding frames (usually, a high level human model is used); and
- **Recognition** – recognizing the behavior, identity, and action of an individual or a group.

A detailed description is given in the next subsections for all these phases with a major focus towards high-level tasks. Each phase topic subsection ends with a discussion subsection that compares the different approaches.

2.1. Model initialization

2.1.1. Main approaches

Initialization of vision-based human motion capture frequently requires the meaning of a humanoid model approximating the appearance, shape, kinematic structure, and beginning posture of the subject to be tracked. Initialization requires prior knowledge of what constitutes an individual. Such knowledge can be separated into categories of Moeslund et al. (2006): kinematic structure, 3D shape, color appearance and body part estimation.

The bulk of vision-based tracking frameworks acquire an initial humanoid kinematic structure incorporating a fix number of joints with specified degrees-of-flexibility. The kinematic initialization is then constrained to the estimation of limb lengths. Commercial marker-based motion capture frameworks normally oblige a fix grouping of movements which separate individual degrees-of-opportunity. Initialization of body pose and limb length from manually identified joint locations using monocular images has been addressed in several works (Barrón & Kakadiaris, 2001, 2003; Parameswaran & Chellappa, 2004; Taylor, 2000). A method for automatically initialization the kinematic structure of the upper body has been investigated by Krahnstöver, Yeasin, and Sharma (2003) and Krahnstoever and Sharma (2004) using motion segmentation of monocular video images. Song, Goncalves, and Perona (2003) presented an unsupervised learning algorithm, which uses point characteristic tracks from jumbled monocular video sequences to automate the process of developing triangulated models of entire body kinematics. Techniques that determine the kinematic structure from 3D shape sequences recreated from different perspectives have also been proposed (Brostow, Essa, Steedly, & Kwatra, 2004; Cheung, Baker, & Kanade, 2003; Chu,

Table 1
Automatic human behavior detection from video keywords by publication year.

Keywords	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
Depth images	0	0	0	1	0	1	2	1	1	0	2	3	2	2	2	17
Stereo cameras	2	1	0	1	0	0	0	0	1	0	0	0	0	0	0	5
Kinect	0	0	0	0	0	0	0	0	0	0	1	3	2	2	1	9
Multiple cameras	0	0	0	1	0	1	2	1	0	2	0	1	1	1	2	12
3D	0	1	1	1	0	0	1	0	3	1	0	3	2	3	3	19
Video dataset	1	0	0	1	1	5	4	2	2	4	1	4	2	3	2	32
Background subtraction	0	0	0	1	0	3	3	2	1	1	1	1	0	1	5	19
Object tracking	0	0	2	1	1	4	4	3	2	2	3	4	1	4	5	36
Occlusion handling	0	0	1	2	0	2	2	2	4	3	3	3	3	2	4	31
Histogram Of Gradients (HOG)	0	0	0	0	0	4	5	3	0	2	2	2	0	0	3	21
Human motion-based features	1	1	0	0	1	3	3	3	1	3	3	2	2	2	3	28
Blobs	0	1	2	1	0	2	1	1	0	1	0	1	0	0	4	14
Neural Network (NN)	2	1	3	1	2	3	0	1	0	1	1	1	0	1	0	17
Support Vector Machine (SVM)	1	1	1	1	1	4	2	2	1	1	2	2	1	1	6	27
Vision	1	1	2	0	0	2	1	1	2	2	1	1	2	2	3	21
Behavior	0	0	1	2	2	4	2	0	2	0	0	1	1	2	5	22
Total	8	7	13	14	8	38	32	22	20	23	20	32	19	26	48	330