Expert Systems With Applications 63 (2016) 97–111

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

A survey on using domain and contextual knowledge for human activity recognition in video streams

Leonardo Onofri[a], Paolo Soda[a,*], Mykola Pechenizkiy[b], Giulio Iannello[a]

[a] Department of Engineering, University Campus Bio-Medico of Rome, Via Alvaro del Portillo 21, 00128 Roma, Italy
[b] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands

ARTICLE INFO

Article history:
Received 16 July 2015
Revised 29 February 2016
Accepted 8 June 2016
Available online 16 June 2016

Keywords:
Human activity recognition
Computer vision
A priori knowledge
Contextual information
Reasoning

ABSTRACT

Human activity recognition has gained an increasing relevance in computer vision and it can be tackled with either non-hierarchical or hierarchical approaches. The former, also known as single-layered approaches, are those that represent and recognize human activities directly from the extracted descriptors, building a model that distinguishes among the activities contained in the training data. The latter represent and recognize human activities in terms of subevents, which are usually recognized my means of single-layered approaches. Alongside of non-hierarchical and hierarchical approaches, we observe that methods incorporating a priori knowledge and context information on the activity are getting growing interest within the community. In this work we refer to this emerging trend in computer vision as knowledge-based human activity recognition with the objective to cover the lack of a summary of these methodologies. More specifically, we survey methods and techniques used in the literature to represent and integrate knowledge and reasoning into the recognition process. We categorize them as statistical approaches, syntactic approaches and description-based approaches. In addition, we further discuss public and private datasets used in this field to promote their use and to enable the interest readers in finding useful resources. This review ends proposing main future research directions in this field.

1. Introduction

Human activity recognition in video streams is an active research area presenting some of the most promising applications of computer vision such as network-based surveillance, content-based video analysis, user-interface and elderly monitoring. Network-based surveillance systems provide interactive, real-time monitoring which increases human efficiency and accuracy, especially with the growing number of cameras (Lin, Sun, Poovandran, & Zhang, 2008; McKenna, 2003; Niu, Long, Han, & Wang, 2004). Content-based video analysis and automatic annotation permit efficient searching, e.g. finding tackles in soccer matches or typical dance moves in music videos (Chang, 2002; Dimitrova, 2003; Hanjalic, Lienhart, Ma, & Smith, 2008). In the user-interface application domain, activity recognition can complement speech recognition and natural language understanding for helping in creating computers that can better interact with humans (Choi, Cho, Han, & Yang, 2008; Pentland, 1998; Shang & Lee, 2011). Finally, monitoring systems which recognize activities of daily living (ADL) can be applied to home care technologies for elderly, reducing the costs and burdens of care-giving while increasing safety and autonomy in old age (Cardinaux, Bhowmik, Abhayaratne, & Hawley, 2011; Khan & Sohn, 2011; Zouba, Boulay, Bremond, & Thonnat, 2008).

The general task of human activity recognition consists in labelling videos that contain human motion with activity classes. To this aim, activity recognition systems cope with a variety of issues, which depend on factors such as the type of acquired videos, the number of persons involved in the activity, the complexity of performed activities and so on. Moreover, these systems could face related topics such as human detection, human movement tracking and person identification, that might be used as lower level modules of an activity recognition system.

The recognition of human activities can be performed at various levels of abstraction. Hence, the goal of an activity recognition system may comprise, for instance, simple movements like “left leg forward” or “arm stretching”; higher complex movements like “running” or “handshaking”; compositions of low-level movements like “jumping hurdles” or “table clearing”. One of the earliest attempt to propose a general definition of human motion was performed by Bobick (1997). He defined a *movement* as the most atomic human motion, an *activity* as a sequence of movements and an *action* as a large-scale event, typically including

* Corresponding author. Fax: +39 06 225419609.
E-mail addresses: leonardonofri@gmail.com (L. Onofri), p.soda@unicampus.it (P. Soda), m.pechenizkiy@tue.nl (M. Pechenizkiy), g.iannello@unicampus.it (G. Iannello).

http://dx.doi.org/10.1016/j.eswa.2016.06.011
0957-4174/

interaction with the environment. Conversely, Turaga, Chellappa, Subrahmanian, and Udrea (2008) defined an *action* as a simple motion pattern usually executed by a single person and typically lasting for short durations of time, whereas an *activity* is a complex sequence of actions performed by several humans who could interact with one another. Moreover, Poppe (2010) adopted a hierarchical scheme as well, defining three levels of abstraction: the lowest level is named the *action primitive*, an *action* is a composition of action primitives that describes a whole-body movement and the *activity* contains a number of actions with a high-level interpretation of the movement.

The aforementioned definitions contain some evident inconsistencies. To avoid any confusion in terminology, we use the term *activity recognition* as the general motion categorization framework, irrespective of the abstraction level actually investigated. When a classification system deals with simple activities that do not show any hierarchy, there is no reason to introduce different definitions and we just use the term activity. On the contrary, when focusing on high level motion understanding, where the approaches typically rely upon a certain degree of hierarchy, following (Aggarwal & Ryoo, 2011) we use the concepts of *event* and *subevents*. A subevent is the lower level movement that is to be recognized, wherein the final goal is the recognition of a higher level activity (the event). For example, we will use the term subevent for the "left leg forward" movement, where the goal is to recognize the event "running", whereas we will use the term subevents for the "running" and "jumping" movements, where the goal is to recognize the event "jumping hurdles". Note that while referring to this kind of high-level activity we will often use the term of composite activities for stressing their property of being characterized by an event composed of subevents.

In a video the information is conveyed in the form of spatio-temporal pixel intensity variations and thus, extracting a suitable set of descriptors is an important prerequisite of any activity recognition system. Once they have been extracted and a set of class labels has been defined, human activity recognition can be formulated as a classification problem that can be tackled with either non-hierarchical or hierarchical approaches (Aggarwal & Ryoo, 2011; Vishwakarma & Agrawal, 2013).

The former, also known as single-layered approaches, are those that represent and recognize human activities directly from the extracted descriptors, building a model which distinguishes among the activities contained in the training data. Single-layered approaches are most effective when a pattern describing an activity can be captured from training sequences; these approaches are suitable for the recognition of gestures and actions, such as relatively simple (and short) sequential movements of humans (e.g., walking, jumping, and waving) (Gorelick, Blank, Shechtman, Irani, & Basri, 2007; Poppe, 2010; Schuldt, Laptev, & Caputo, 2004).

The latter represent and recognize human activities in terms of subevents, which are usually recognized my means of single-layered approaches. Hierarchical methodologies are able to recognize high-level activities because of their ability to incorporate knowledge on the activity structure, making the recognition process conceptually understandable and computationally tractable.

Alongside of non-hierarchical and hierarchical approaches, we observe that methods incorporating *a priori knowledge* and *context information* on the activity (see Section 2 for their definition) are getting growing interest in the literature. In this work we refer to this emerging trend in computer vision as *knowledge-based human activity recognition* (KBAR) with the objective to cover the lack of a summary of these methodologies. More specifically, we survey methods and techniques used to represent and to integrate knowledge and reasoning into the recognition process, whereas we do not focus on low-level modules such as body structure analysis, tracking and feature extraction.

The paper is organized as follows: Section 2 discusses the exploitable knowledge, whereas Section 3 overviews the approaches for knowledge-based exploitation in human activity recognition. Section 4 present the available datasets for testing the methodologies. Section 5 discusses the surveyed contributions, whereas Section 6 provides future directions and concludes the paper.

1.1. Comparisons with previous reviews

Previous reviews on human activity recognition have focused on different aspects of motion understanding. Bobick (1997) described different approaches dividing his analysis in three different levels of abstraction, i.e. movements, activities and actions. Aggarwal and Cai (1999) and Wang, Hu, and Tan (2003) discussed body structure analysis, tracking and recognition. Kruger, Kragic, Ude, and Geib (2007) reviewed human action recognition approaches while classifying them on the basis of the complexity of features involved in the action recognition process. Their reviews focused especially on the planning aspect of human action recognitions, considering their potential application to robotics. Poppe (2010) considered image representation and video classification, limiting his survey to simple activity recognition. Turaga et al. (2008) and Aggarwal and Ryoo (2011) focused on both simple and complex human activities, describing different approaches in terms of feature extraction and classification algorithms. In their paper, approaches are categorized on the basis of the complexity of the activities and in terms of the recognition methodologies they use. Vishwakarma and Agrawal (2013) and Suriani, Hussain, and Zulkifley (2013) directed their surveys towards surveillance systems. The former offers a summary for activity recognition in video surveillance, integrating the surveyed papers presented in Aggarwal and Ryoo (2011), and providing a discussion on object tracking. The latter focused on frameworks used in sudden event recognition, defined as a subset of an abnormal event in video surveillance applications, reporting also the requirements and a comparative studies of a sudden event recognition system. Recently, Ziaeefard and Bergevin (2015) surveyed methodologies for activity recognition in still images and videos using semantic features. The review identifies the pose, the poselet, the objects, the scene, and the attributes as semantic features and it mostly discusses how they can be extracted and used to recognize the human activities. It mentions that hierarchical representation and reasoning mechanisms can be used to recognize the activities, and it briefly discusses potential applications where semantic approaches may be of assistance. Nevertheless, this work does not address how knowledge needed to exploit semantic information can be represented and integrated into the recognition process.

2. Exploitable knowledge

Knowledge exploitation is an established approach in the data mining literature, since it is helpful for selecting suitable classification techniques, pruning the space of hypothesis and improving the overall performance (Nigro, Císaro, & Xodo, 2008). The several advantages of knowledge exploitation can be summarized as follows (Crevier & Lepage, 1997):

- With an explicit knowledge arrangement, data contradictions and omissions become apparent, thus suggesting alternative means of extracting information from videos and images.
- Knowledge-based techniques permit to design and develop in an intuitive (visual) manner the recognition system and to extract information from examples.
- Explicit knowledge representation allows the separate description and the parallel use of knowledge pertaining to different domains, such as knowledge about image processing, knowl-