Pattern Recognition 48 (2015) 2329–2345

Contents lists available at ScienceDirect

Pattern Recognition

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

Semantic human activity recognition: A literature review

Maryam Ziaeefard*, Robert Bergevin

Computer Vision and Systems Laboratory, Department of Electrical and Computer Engineering, Laval University, Quebec, Canada G1V0A6

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form
7 February 2015
Accepted 6 March 2015
Available online 14 March 2015

Keywords:
Human activity recognition
Pose
Poselet
Attribute
Human-object interaction
Scene
Survey

ABSTRACT

This paper presents an overview of state-of-the-art methods in activity recognition using semantic features. Unlike low-level features, semantic features describe inherent characteristics of activities. Therefore, semantics make the recognition task more reliable especially when the same actions look visually different due to the variety of action executions. We define a semantic space including the most popular semantic features of an action namely the human body (pose and poselet), attributes, related objects, and scene context. We present methods exploiting these semantic features to recognize activities from still images and video data as well as four groups of activities: atomic actions, people interactions, human–object interactions, and group activities. Furthermore, we provide potential applications of semantic approaches along with directions for future research.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Human activity recognition is being leveraged for an increasingly wide variety of computer vision applications. What all of these works have in common is to study some aspects of human–computer interaction. Recognizing activities can range from a single person action to multi-people activity recognition. Generally, an action is defined as a single person activity but we use the terms action and activity interchangeably.

A number of surveys have been published in activity recognition during the last decade. Most of the earlier reviews have focused on the introduction and general summarization of activity recognition methodologies [1–3]. A study by Turaga et al. [4] covered human activity recognition methods with a categorization based on the complexity of activities and recognition methodologies. Various challenges in action recognition were addressed and limitations of different approaches were discussed in [5]. Recently, Aggarwal and Ryoo [6] conducted a survey emphasizing activity recognition methods for four groups of activities (atomic action, people interaction, human–object interaction, and group activity). They classified activity recognition methodologies into two categories: single-layered approaches and hierarchical approaches. Single-layered methods represent and recognize human activities directly based on sequences of images. On the other hand, hierarchical approaches describe high-level human activities by using simpler activities called sub-events which are suitable for the analysis of complex activities. Aggarwal and Ryoo [6] also mentioned a few semantic approaches without clearly explaining what semantics is and why it should be used. In this survey, we aim to cover the methods in the literature which address semantic activity understanding.

Human activity recognition methods can also be classified according to their input data. Traditional action recognition approaches used videos or image sequences while recent studies started to explore action recognition in still images. Compared to the video-based action recognition, still image-based action recognition has some special properties. For example, there is no motion in a still image, and thus many spatio-temporal features and methods that were developed for traditional video-based action recognition are not applicable to still images. A recent survey [7] presents a detailed overview of the existing approaches in still image-based action recognition and explains various features as well as related databases which have been used in analyzing actions in still images.

Different levels of features have been used in activity recognition methods. Traditional action recognition methods rely mostly on tracking, and motion capture. Mid-level features such as spatio-temporal and bag-of-word features are used by recent approaches. Semantic features, meanwhile, are aimed to answer questions such as "what does it mean to do an action?" or "How do we understand an action?". The term semantics refers to the study of meaning. For example, it is meaningful that a car and road appear in the same images, while a giraffe and a kitchen should not. A detailed definition of this term will be provided in Section 2.

Semantic features are useful to address the problem of intra-class variability. Intra-class variability refers to the differences in the same group of actions and how different instances of the same

* Corresponding author. Tel.: +1 418 656 2131; fax: +1 418 656 3159.
E-mail address: maryam.ziaeefard.1@ulaval.ca (M. Ziaeefard).

http://dx.doi.org/10.1016/j.patcog.2015.03.006
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

action resemble each other. As shown in Fig. 1, people may perform the same action in different ways or even the same person may perform one action differently in different situations. In addition, humans vary significantly in appearance due to changes in clothing, body shape and viewpoint. Semantic features help to distinguish similar actions that differ visually but have common semantics.

Semantic approaches apply the human understanding of the activity. The human ability to recognize actions does not rely only on visual analysis of human body postures but also requires additional sources of information such as context or scene, knowledge about objects related to activities, or knowledge about the visual characteristics of activities. On the other hand, non-semantic approaches, here, refer to methods representing actions *only* in some form of low-level features such as silhouette, gradients, and optical flow. They do not incorporate human knowledge about activities. Non-semantic approaches capture the appearance and motion characteristics while semantic approaches describe inherent characteristics of activities. Non-semantic approaches are ideally appropriate for simple actions. However, they fail in complex situations due to the lack of semantics they represent.

To classify semantic approaches, we introduce a feature space called the "semantic space" which includes human knowledge about activities such as the body part (pose and poselet), object, scene, and attribute features. The semantic space is illustrated schematically in Fig. 2. Based on exploiting these features, we categorize semantic methods into three categories: methods based on body parts, methods based on objects/scenes, and methods based on attributes.

The first feature of the semantic space is the body part. Neuropsychological studies indicate that semantic knowledge of human body parts might be distinct from knowledge of other object categories. Downing et al. [8] identified a subpart of the human extrastriate cortex involved in the visual processing of the human body and body parts, namely extrastriate body area or EBA. Their experimental results reveal that the EBA responds strongly and selectively to a variety of pictures of human bodies and body parts. The EBA may be crucial for perceiving the position and configuration of one's body, possibly as part of a general system for inferring the actions and intentions of others. Also, EBA may be involved in perceiving the configuration of one's own body. Peelen and Downing [9] and Schwarzlose et al. [10] worked also on the body selectivity of the brain. Methods for pose-based action recognition can either use pose estimation results as an input for the action recognition step or address both pose estimation and action recognition concurrently. The latter approach has the advantage that errors due to inaccurate pose estimation will have less of an effect on the final quality of activity recognition. Semantics also captures salient body parts during an action which is referred to as a poselet. In 2D/3D images,

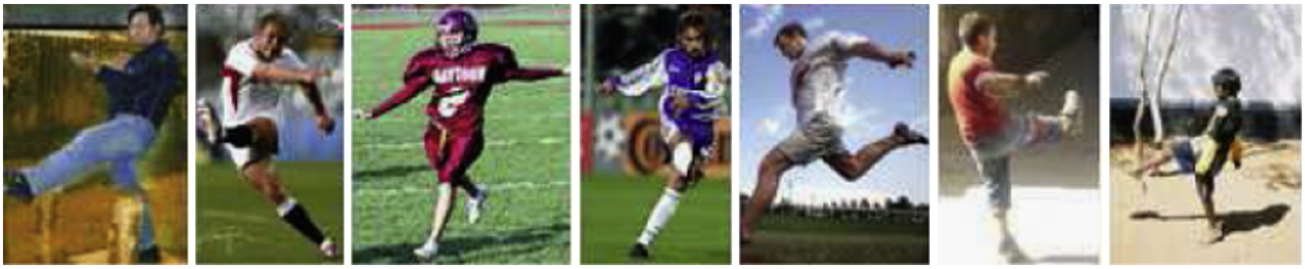

Fig. 1. "Kicking" action. The same actions appear different due to different camera angles, clothes, body shapes, etc.

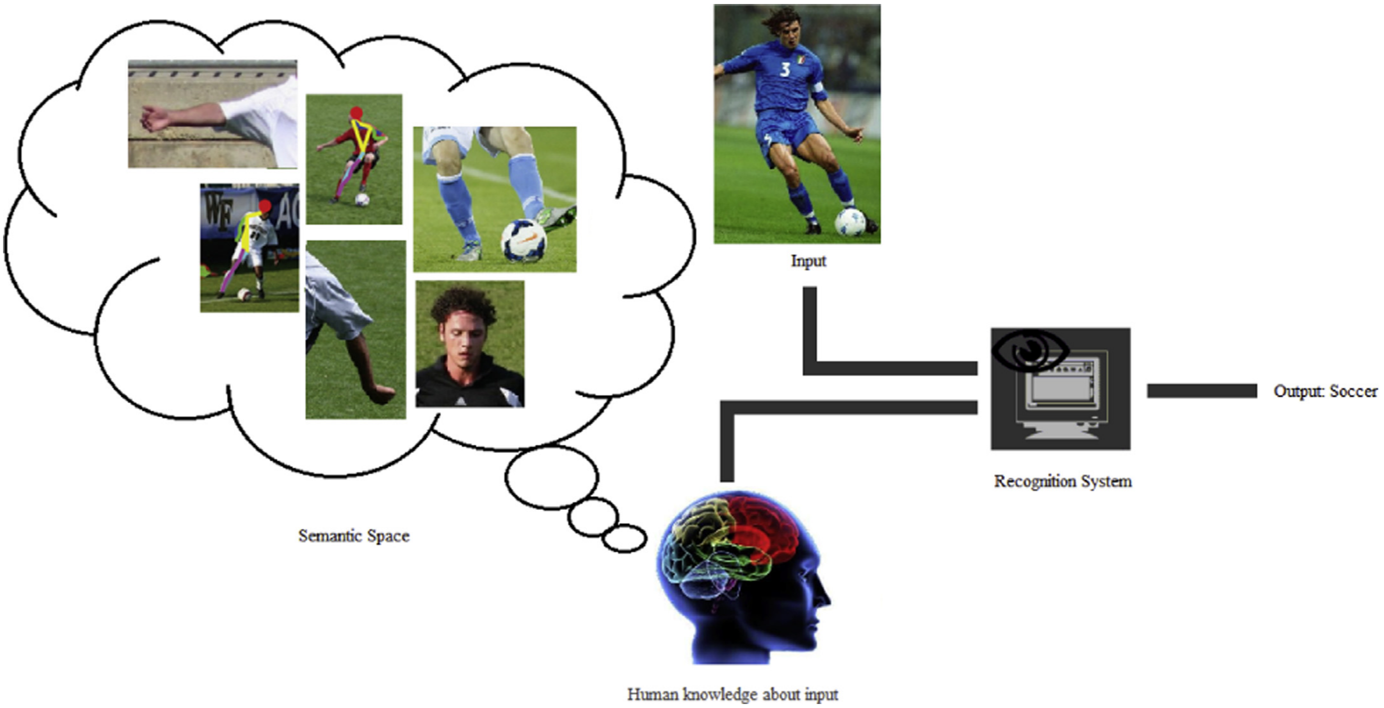

Fig. 2. *Semantic space*: observing an action, e.g. "playing soccer", the human uses his knowledge to recognize the activity. We define a semantic space containing pose (specific body pose in soccer), poselet (extended right arm, straight left arm), object (soccer ball, interaction between one leg and a soccer ball), scene (soccer field), and attribute (looking-down head) which are illustrated in the figure.

Download English Version:

https://daneshyari.com/en/article/10361268

Download Persian Version:

https://daneshyari.com/article/10361268

Daneshyari.com