J. Vis. Commun. Image R. xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Pose-based human action recognition via sparse representation in dissimilarity space

Ilias Theodorakopoulos *, Dimitris Kastaniotis, George Economou, Spiros Fotopoulos

Electronics Laboratory, Department of Physics, University of Patras, Patras 26500, Greece

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Action recognition
Sparse representation
Dissimilarity representation
Pose representation
Articulated human motion
RGB-D sensors
Angular features
Pose encoding

ABSTRACT

Human actions can be considered as a sequence of body poses over time, usually represented by coordinates corresponding to human skeleton models. Recently, a variety of low-cost devices have been released, able to produce markerless real time pose estimation. Nevertheless, limitations of the incorporated RGB-D sensors can produce inaccuracies, necessitating the utilization of alternative representation and classification schemes in order to boost performance. In this context, we propose a method for action recognition where skeletal data are initially processed in order to obtain robust and invariant pose representations and then vectors of dissimilarities to a set of prototype actions are computed. The task of recognition is performed in the dissimilarity space using sparse representation. A new publicly available dataset is introduced in this paper, created for evaluation purposes. The proposed method was also evaluated on other public datasets, and the results are compared to those of similar methods.

1. Introduction

Human motion analysis has recently emerged as one of the most dynamic research topics in the field of computer vision. Especially, the task of assigning labels of action classes to data sequences induced by human activity has attracted much of attention during the past two decades. The reason is that applications in various domains such as surveillance, security, assisted living, and others would greatly benefit from the incorporation of algorithms able to model and recognize ongoing human activities.

Several taxonomies of human activity have been proposed. Aggarwal et al. [1] adopt the categorization of activities in gestures (elementary movements of body parts), actions (single person activities), interactions (actions involving two or more persons or objects) and group activities, based on their complexity. Moeslund et al. in [20] proposed a hierarchy based on distinction between action primitives, which are atomic movements possibly in limb level, actions and activities. Actions consist of several action primitives and describe more complex, whole-body movements, while activities are considered as a set of actions, sequentially executed, bearing conceptual information about the purpose of their execution. In this paper we focus on single-person actions, as outlined by both of the above taxonomies, without explicitly considering interactions with objects of the environment.

From the data perspective, human activity is usually captured in videos. Thus, the vast majority of the proposed techniques in the research area of vision-based human activity analysis are designed to handle sequences of images illustrating consecutive instances of human motion. Several surveys of this field exist in the literature, offering different categorizations of the preceding published methods. In [1] activity recognition methodologies are classified into single-layered, where activities are recognized directly from the sequence of images, and hierarchical approaches, where each higher-order activity is considered as an aggregation of simpler actions called sub-events. Single-layered are furthered classified into space–time approaches, where the input video is utilized as a three-dimensional volume, and sequential approaches where the video serves as a sequence of observations. Hierarchical methods are also sub-divided into statistical, syntactic and description based methods.

Poppe in [24] categorizes methods for recognition of human actions based on the features extracted from the image sequences. Under the proposed taxonomy methods are divided into global representations (grid-based and space–time volume based), local representations (local descriptors, correlation-based, space–time points PoIs) and application specific representations classes. A further sub-division is based on the employed classification strategy, distinguishing between temporal state-space models and direct classification approaches. Other taxonomies are based on the distinction between 2D, 3D and recognition methods [14] and functional characteristics such as initialization, tracking, pose estimation and recognition.

In recent years the trends in the field of action recognition have been shifted towards utilization of abstract, low-level features stemming directly from the raw video data, in order to deal with

* Corresponding author.
E-mail address: iltheodorako@upatras.gr (I. Theodorakopoulos).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.03.008

the demand for analysis of natural unconstrained videos. Early work though [17], showed that the information of joint positions over time, is sufficient for humans to distinguish actions. Pose-based approaches for action recognition are based upon this observation, considering human actions as a sequence of articulated poses. However, pose estimation requires accurate tracking of body parts, which is known to be a very challenging problem considering the non-rigidness and self-occlusion of human body. Traditionally, pose estimation was performed using data captured by specialized devices requiring electromagnetic markers attached to the body and fully controlled environment. These devices were both expensive and obstructive, and so the practical applications were limited. Recently though, several low-cost devices have been released such as Microsoft's Kinect or ASUS's Xtion Pro Live, enabling markerless pose estimation with reasonable accuracy, using Vision-based pose estimation techniques. This type of techniques attempt to provide a markerless solution to human pose-based motion analysis problems, incorporating data from RGB or RGB-D cameras. In [25] Poppe provided taxonomy of this type of methods, classifying the approaches into model-based and model-free, depending on whether an a priori body model is employed or not. Analysis of model-based methods is further divided into modeling (body models, image descriptors, camera and environmental considerations) and estimation phase (Top-down and bottom-up estimation, motion priors, single-multiple hypothesis tracking).

Despite the demanding initial processing though, pose-based approaches exhibit many advantages. Firstly, pose is a higher-order feature which embodies much of the action-relevant information and thus, learning can be much simpler comparing to approaches using low-level characteristics. Secondly, intra-class variations of actions stemming from anthropomorphic differences between individuals are expressed mildly on pose data. Additionally, pose representations such as 3D skeleton, which is the representation of choice in the current work, exhibit some desirable characteristics such as viewpoint invariance. The above features are of major importance to the performance of this type of systems, since the aim of recognition is the appropriate generalization over the action variations appearing in the training set, as opposed to e.g. gait recognition or other human action-based biometric applications. Recently, Yao et al. in [32,33] raised the question of whether it is useful to perform pose estimation for the action recognition task or it is better to extract low-level appearance features directly from video data. Their experiments indicate that pose-based approach outperforms the appearance-based on the same dataset, using the same classifier. They further showed that even with high levels of noise the pose-based features either match or outperform the appearance-based, and thus high precision in pose estimation seems to be unnecessary.

To this direction, several pose estimation algorithms have been proposed, able to produce skeletal pose representation in real-time, incorporating depth data from commercial RGB-D sensors. Such an approach is the state-of-the-art algorithm of Shotton et al. [27] which is utilized in the current work, implemented in the Microsoft Kinect SDK. Despite that the implemented algorithm is quite sophisticated, the inferred noise from the depth sensor of such low-cost commercial devices is significant. Additionally, the self occlusion of human body under the conditions of monocular view is inevitable. Various approximations though, have to be made internally, in order to extract a pose estimation in real-time, consistent with the real-world constraints of the human physiology and motion. Although these approximations result a robust behavior in gamming applications, which is the main target for such devices, they introduce significant error to the calculated joint positions relative to the actual, as opposed to expensive and invasive MoCap systems. This leads classical features related to motion analysis, such as 3D joint trajectories, joint angles, simple relational features etc., derived from such data, to exhibit poor performance in terms of activity recognition. Therefore, alternative approaches have to be considered both in feature and learning level, in order to boost the performance of systems employing such devices.

In this paper we propose a method for pose-based human action recognition, able to be incorporated with the low-cost devices for markerless pose estimation, available in the market. In the proposed scheme, 3D coordinates of joints from a skeleton model are considered as inputs. Skeletal data are initially processed so as to obtain robust and invariant pose representations. In order to boost the recognition performance, the DS-SRC classification scheme proposed in [28] is incorporated. According to this approach, actions are initially represented by vectors of dissimilarities to a set of prototype actions. Then, the task of recognition is performed into the dissimilarity space using sparse representation-based classification. We also introduce the publicly available UPCV Action dataset which was created in order to experimentally evaluate the proposed method. In order to highlight the essential role of DS-SRC scheme regarding the performance of the proposed scheme, we compare to the performance of a standard k-NN classifier directly applied on the dissimilarity data, and to a scheme proposed by Duin et al. in [8] consisting of SVM classification into the dissimilarity space. Also we evaluate the proposed method on other public datasets in order to compare to the performance of similar techniques from the literature.

The rest of this paper is organized as follows: The details on the processing of the skeletal data along with the proposed representation are described in Section 2. The incorporated DS-SRC classification scheme is discussed in two sections: First, in Section 3 the dissimilarity representation of human actions is formulated, and in Section 4 the incorporated sparse representation-based classification scheme is described. A description of an introduced dataset, used during evaluation is given in Section 5. The evaluation procedure is described in Section 6 and conclusions are drawn in Section 7.

2. Pose representation

As mentioned in the introduction of this paper, skeletal representation of human poses is considered as input to the proposed method. This type of representation consists of coordinates of joints in the 3-dimensional space, and a wireframe skeleton model. In Fig. 1 an example of the incorporated skeleton model is illustrated, consisting of 20 points corresponding to the joints of a basic human body model.

Starting with a set of joint coordinates in a three dimensional space, the goal is to extract a set of features in order to represent

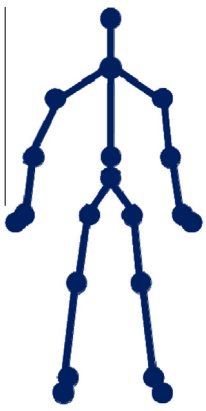

Fig. 1. Example of skeletal model of 20 joints.

Please cite this article in press as: I. Theodorakopoulos et al., Pose-based human action recognition via sparse representation in dissimilarity space, J. Vis. Commun. (2013), http://dx.doi.org/10.1016/j.jvcir.2013.03.008