Pattern Recognition 47 (2014) 238–247

Contents lists available at ScienceDirect

Pattern Recognition

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

Ongoing human action recognition with motion capture

Mathieu Barnachon [a,*], Saïda Bouakaz [a], Boubakeur Boufama [b], Erwan Guillou [a]

[a] Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
[b] School of Computer Science, University of Windsor, Windsor, ON, Canada N9B 3P4

ARTICLE INFO

Article history:
Received 25 July 2012
Received in revised form
9 May 2013
Accepted 11 June 2013
Available online 26 June 2013

Keywords:
Human action recognition
Histogram
Ongoing recognition
Human–computer interaction

ABSTRACT

Ongoing human action recognition is a challenging problem that has many applications, such as video surveillance, patient monitoring, human–computer interaction, etc. This paper presents a novel framework for recognizing streamed actions using Motion Capture (MoCap) data. Unlike the after-the-fact classification of completed activities, this work aims at achieving early recognition of ongoing activities. The proposed method is time efficient as it is based on histograms of action poses, extracted from MoCap data, that are computed according to Hausdorff distance. The histograms are then compared with the Bhattacharyya distance and warped by a dynamic time warping process to achieve their optimal alignment. This process, implemented by our dynamic programming-based solution, has the advantage of allowing some stretching flexibility to accommodate for possible action length changes. We have shown the success and effectiveness of our solution by testing it on large datasets and comparing it with several state-of-the-art methods. In particular, we were able to achieve excellent recognition rates that have outperformed many well known methods.

1. Introduction

Human action recognition is a challenging research problem in computer vision that, if solved, would enhance numerous applications in areas ranging from Human Computer Interface (HCI) to entertainment. For instance, human action recognition could help to identify suspicious activities. In entertainment applications, recognizing players' actions makes the game more attractive, more user-friendly and increases its potential. Given its importance in numerous applications, the problem of action recognition has attracted a great deal of research works over the last decades. Although ambiguous 2D images have been traditionally used as inputs for action recognition, numerous researchers have started using MoCap data for action recognition [30,44,40]. In particular, newly available low cost depth sensors, such as Microsoft Kinect and its real-time MoCap system [35], can be used to enhance the user's experience with games, serious games, presentation softwares, etc.

This paper proposes a novel examplar-based human action recognition system, that uses MoCap poses as input. We first extend the concept of histograms to the case of poses. Then, spatio-temporal series of poses are clustered to create a statistical representation of actions. Note that when an action consists of cycles, for example walking, its histogram, or part of it, is affected by a scale only. We have also introduced an incremental and memory efficient structure, the integral histogram, to allow for ongoing activity recognition. Finally, a dynamic programming algorithm, inspired from the Dynamic Time Warping (DTW) method [33], is used to compare sub-actions and to compute the recognition score between multiple human action instances. To the best of our knowledge, this paper is the first to propose a solution using histograms of 3D Motion Capture data for action recognition. The efficient formulation of histograms has made it possible to learn and recognize actions during their progress. We have validated our proposed approach with extensive tests on well-known benchmark datasets, and we have compared it to several state-of-the-art methods. The obtained results have clearly shown the success and effectiveness of our solution, even in the presence of noise and/or similar actions in the datasets.

2. Previous works

This section summarizes major previous works in human action recognition, and briefly surveys three related issues: the body skeletonization, the body shape analysis and the extraction of feature points. For an extensive survey on human action recognition, the interested reader may consult [1].

The skeleton is usually easy to extract and is known to make an efficient and compact representation of a shape, like the human body [39]. The first body skeletonization method to analyze actions was proposed by Fujiyoshi et al. [12]. Their method performs a skeletonization of the body contour to identify walking,

* Corresponding author. Tel.: +33 426234445.
E-mail addresses: mathieu.barnachon@gmail.com, mathieu.barnachon@liris.cnrs.fr (M. Barnachon), boufama@uwindsor.ca (B. Boufama).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.06.020

running and gait. Their solution is simple to use and depends only on a single 2D image to extract the skeleton. Although the recognition process of this method is very efficient for simple activities, it suffers from the simple "star" skeletonization problem as well as from visual ambiguities. Ziaeefard and Ebrahimnezhad [47] have proposed an improvement for this method. They have introduced a normalized-polar histogram, obtained from the extracted "star" skeleton, that corresponds to a cumulative skeleton during one action cycle. In particular, they have analyzed different skeletonization methods and proposed an SVM classification technique to recognize actions. Tran et al. [37] have used a different skeletonization method and have achieved better results on the same datasets. However, as they have used polar histograms instead of time-based histogram, the temporal information is lost. Lv and Nevatia [22] have proposed a MoCap-based solution where actions are modeled by a set of virtual key-poses. This is somehow similar to the animation key-poses that represent important poses and/or transitions between sub-parts of actions. This solution is limited by the number of extracted key-poses and by their computational complexity.

Cuntoor et al. [9] have suggested that trajectories contain the most discriminative information that is relevant to human action analysis. Inspired by this observation, Li and Fukui [20] have proposed a trajectory-based solution using Motion Capture data to identify human actions. However, they have only tested their solution on simple cases and not on real human data variations. Using a large database, Han et al. [13] have exploited the skeleton hierarchy to compute trajectories, where actions were represented in a manifold space. As they have used not all but a subset of joints, they needed very large samples in the training set and a high intra-class variation. Therefore, the clustering process of similar actions in their approach was complex and time consuming. Baak et al. [3] and Müller et al. [27] have addressed the problem of action recognition using the idea of Motion Template. They extract patterns from a sequence of animation to recognize actions, transforming the recognition problem into a tractable pattern recognition problem. In [3], a method was proposed to improve MoCap extraction, using a database of priors such as, feet on the ground during the walk, etc.

Recently available, cheap and easy to use, depth sensors have opened new perspectives for solving the problem of action recognition. Raptis et al. [30] have used joint angles as features to recognize dance actions in a game-based application. As mentioned by the authors, their method is limited by the number of actions. In particular, when the number of different classes is large, their error rate increases drastically. With similar input data, Wang et al. [40] have introduced the concept of *actionlet*. They cluster joints and depth neighborhoods in order to be more discriminant. For instance, "drinking from a cup" and "eating a peanut" can be discriminated according to the depth data around the hand. However, the MoCap data has to include depth information, which is not always possible. Related works from the animation community are also relevant to action recognition. Barbič et al. [4] have used the Principal Component Analysis (PCA) to extract a known number of similar behaviors. Beaudoin et al. [5] extracted subsequences of animations, that are similar in the proposed "motion space", then used a graph-based solution to create smooth transitions between animations. Given that these solutions were designed to be animation tools, i.e. to produce smooth transitions between animations, they suffer from the lack of efficient interpretation structures.

Instead of using skeleton, many researchers have used human shape analysis, mostly silhouettes, to address human action recognition. Bobick and Davis [7] have introduced Motion Templates from Motion History Image (MHI), where the recognition problem is turned into a matching problem. Although their system is faster than classical machine learning approaches, it is still time consuming and not flexible enough for extending the database. To address the efficiency of the database, Elgammal et al. [11] have used the "examplar" paradigm with silhouettes. In particular, the Markov model and the "examplar" paradigm lead to a light training database. Although their solution is efficient for adding new actions, it suffers from the view dependency problem, which is inherent to silhouettes. The proposed solution is more appropriate for "simple" gesture recognition than "complex" action recognition. Huang and Trivedi [16] have presented the concept of cylindrical histogram, where multiple views are used to construct a 3D histogram of voxels. Weinland et al. [41] have adapted the 3D histogram process to the examplar paradigm for view-independent learning from multiple views. Boulgoris and Chi [8] have proposed a hybrid solution that uses labelled body parts from silhouettes. Even though their solution is efficient for gait analysis, its use for general action recognition is hindered by its labeling process that has to be done separately. Xiong and Liu [43] have also used a Markov model with silhouettes to recognize mainly simple behaviors. Yilmaz and Shah [46] considers a silhouette as a 2D surface and construct a 3D surface from a sequence of spatio-temporal silhouettes. Then, they extract interest points from the obtained 3D surface, creating something like a trace of an action. Unfortunately, their process is not real-time because it requires an expensive stage of silhouette correspondence for computing the 3D surface. In addition, the obtained lengthy volume is dependent on the silhouette quality, which is prone to errors. Ahmad and Lee [2] have proposed an extension of the MHI where they have used an SVM to cluster actions. As they were using too many parameters in their system, it was difficult to draw a strong conclusion from their results. Tseng et al. [38] have developed a silhouette-based approach, where silhouettes were used as characteristic vectors. Their actions were clustered using a dimension reduction method, then the *k*-nearest-neighbor algorithm was used on a temporal graph in the recognition stage. Because their solution depends on the quality of the extracted silhouettes, the recognition success might suffer from it.

Many other previous research works on action recognition have used feature points in a spatio-temporal framework. Laptev and Lindeberg [18] extended the Harris and Stephen detector [14] to the spatio-temporal case. Dollar et al. [10] have proposed another spatio-temporal feature detector, especially designed for cyclic motion in actions. They introduced the concept of cuboid, widely followed by others, where each cuboid encodes information about a local neighborhood. As the spatial locations of cuboids were ignored, it has lead to the concept of bag of words. Ryoo [31] has used these bags of words to construct histograms and use them for ongoing activity recognition. Their solution uses 2D image features, instead of our 3D primitives, and is not examplar-based. In particular, their training phase requires more processing, making it difficult to add new actions. One can also consider the work of Scovanner et al. [34] where the SIFT detector was extended to the 3D case of action recognition. Their solution is usually used with an extrusion of spatio-temporal volume from 2D images, a complex process that is also sensitive to the background extraction result.

More recently, many researchers have worked on spatial configurations. Wong et al. [29] have proposed an extension to the pLSA space to model spatial relations [42]. Ryoo and Aggarwal [32] introduced the *Spatio-Temporal Relationship match* (STR match) to consider the spatial information with the temporal one. Yao et al. [45] obtained a non-linear latent space to discriminate between complex activities in a kitchen. Although their solution can be considered efficient, their latent space is complex to compute and need a huge training set to be effective. By contrast, our solution can work with smaller training sets. In particular, our proposed method outperformed their recognition rate for the kitchen scene, as shown in the experimental results.