Pattern Recognition Letters 000 (2016) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Human action recognition with skeleton induced discriminative approximate rigid part model☆

Yu Zhou, Anlong Ming*

School of Computer Science, Institute of Sensing Technology and Business (WuXi), Beijing University of Posts and Telecommunications, Beijing, PR China

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Human Action Recognition
Joint
Surface
Part Model

ABSTRACT

Human action recognition has a long research history. Despite various approaches have been designed in the last decades, it still remains challenging in computer vision and pattern recognition. In this paper, we present a skeleton induced discriminative approximate rigid part model for human action recognition, which not only captures the geometrical structure of human body, but also takes rich human body surface cues into consideration. In conventional approaches, the joint feature and the surface feature are discussed separately. While, in our proposed approach, the structural information is embedded into the surface model. In addition, to separate different approximate rigid parts generated from different human activities, a novel sparsity induced feature selection scheme is introduced. This scheme produces a discriminative feature subspace that can best separate different action classes. The presented approach is validated on two widely adopted benchmark datasets, *i.e.* the MSR Daily Activity 3D Dataset and the MSR Action 3D dataset. Experimental results demonstrate the effectiveness of our approach.

1. Introduction

Recognizing the human activity in the cluttered indoor environments is a critical issue in computer vision and pattern recognition. It has many practical applications, like the human-robot interfaces [36], video surveillance, etc. Moreover, other applications of computer vision can also benefit from accurate recognition of human action, e.g. object tracking [35], event detection and recognition. Although several contributed approaches have been introduced in the literatures, it still remains a challenging task. The critical issues include the intra-class variations, such as human pose variation, deformation and self-occlusion, and the extra-class noises, e.g., different actions may have similar appearances in practice.

Most early attempts on action recognition mainly work on color videos [6,8,24]. In these approaches, invariant key points are frequently employed as local features to capture the action of the target. However, the information supplied only by color videos is often insufficient to recognize the action of the human accurately in practice. Recently, the cost-effective depth cameras, like the Kinect RGB-D sensor[1], have attracted much attention from the community. Such cameras provide the 3D depth cue of scene, and hence the activity recognition can naturally benefit from the additional depth information.

In the depth based action recognition, the powerful 3D joint position of the human skeleton can be easily obtained [20]. Features based on the 3D joints of the skeleton are frequently employed to extract the invariant characteristics of human. For example, the method proposed in Wang et al. [26] utilizes the spatial distance between pairwise joints and the local occupancy pattern feature as the feature representations. Since the skeletal features exploit rough structural information of the target, it is capable to address the non-rigid deformation caused by flexible human motions. Dense 3D points sensed by the Kinect sensor further supply accurate depth information. In other words, the geometrical surface can be very informative for recognition.

Certainly, the skeleton cue and the surface cue are complementary to each other. However, these two kinds of feature representations are rarely discussed jointly in the literature. This paper proposes a novel skeleton induced discriminative approximate rigid part model to simultaneously consider these two features in a unified framework for human action recognition. The flow chart of our approach is shown in Fig. 1(a) is the input depth sequence captured by the Kinect sensor; (b) is the surface of the human body; and (c) is the joint (indicated by red dot) and the skeleton (indicated by blue line) of the human captured by Shotton et al. [20]. Instead of generating the feature representation directly by the skeleton or the surface of the human body, in each frame, the whole human body is decomposed into several local approximate

☆ This paper has been recommended for acceptance by Xiang Bai.

* Corresponding author.
E-mail address: minganlong@bupt.edu.cn, mal@bupt.edu.cn (A. Ming).

[1] http://www.xbox.com/en-US/Kinect.

http://dx.doi.org/10.1016/j.patrec.2016.07.025
0167-8655/

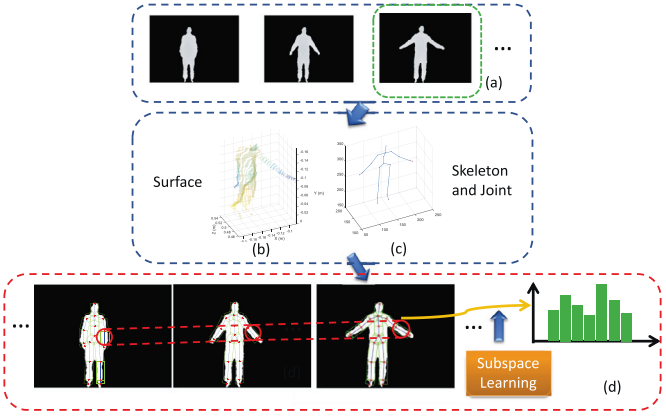

Fig. 1. Flow chart of our approach. (a) the input depth video sequence; (b) the surface of a target; (c) the skeleton of a target; (d) discriminative approximate rigid part model.

rigid parts (LARPs) as shown in (d), which are induced by the human skeleton. For the quantitative representation, Yang and Tian [32] introduce the polynormal, which is a set of surface points expressed by the surface normal, and these points are extracted inside a local space-time subvolume. In our approach, the polynormal is employed to exploit the surface characteristics of each LARP, and the corresponding LARPs in consecutive frames form a global approximate rigid part (GARP) model (highlighted with the red circles in Fig. 1(d)). In addition, the temporal variation of the skeleton is utilized to capture the temporal geometrical structure of the GARP. Both the geometrical structure of the human body and the rich human body surface cues are taken into consideration in our approach. Our model is expected to be more discriminative than existing skeleton or surface based approaches. Another problem should be concerned is that the representation of the GARP is rich yet redundant. To address this issue, a sparsity induced feature subspace learning approach is introduced to produce a discriminative global approximate rigid part (DGARP) model, aiming to separate different action classes.

The main contribution of this paper can be summarized in the following aspects:

- We present a novel discriminative approximate rigid part model, which is induced by the skeleton. The presented approach takes into account both the geometrical structure of the human body and the rich human body surface cues simultaneously.
- A novel sparsity induced feature selection approach is introduced, which produces a discriminative feature subspace for best separating different action classes.
- To demonstrate the efficacy of the presented approach and show its promising performance in comparison with other approaches, experiments are conducted on the widely adopted MSR Daily Activity 3D dataset and MSR Action 3D dataset.

The remainder of this paper is organized as follows: Section 2 briefly reviews related works. Section 3 formally formulates the problem. In Section 4, we describe the discriminative approximate rigid part model. Section 5 reports the experimental results on two famous human action recognition datasets. Conclusion and future work are given in Section 6.

2. Related work

Action recognition has a long research history, a detailed survey can be found in Aggarwal and Ryoo [9]. The skeleton based representations have been well studied in the 2D shape, e.g., [2,18]. Since the reliable 3D structure, i.e., the 3D skeleton, of the human can be easily obtained by the cost-effective depth sensor [19], it attracts much attention for the human action recognition, e.g., Lv and Nevatia [13] utilize the Hidden Markov model to reflect the transition probability of the 3D joints. Han et al. [7] employ the conditional random field to model the 3D joint positions. Wang et al. [26] present a novel joint based image feature to represent the action. In general, the joint or skeleton based approaches well capture the geometrical structure of the human action, but it is insufficient to express the activity of the human without the appearance cues.

The 2D contour based representations are also well discussed in the existing approaches, e.g., [1,3,14,15]. In the early research literatures, reliable key point based strategies are frequently utilized, e.g., [8]. And then the motion trajectory based approaches are introduced, e.g., [24,29]. Furthermore, the holistic approaches become popular in recent years, e.g., Yang et al. [33] summarize the whole depth sequences into a motion map, and the HOG feature [5] is extracted from the motion map, which is employed to reflect the whole sequence. Vieira et al. [22] divide the whole depth sequences into several spatio-temporal grids, and the global occupancy patterns are employed to represent the whole sequence. Oreifej and Liu [17] introduce the HON4D feature, which captures the change rate of the surface in a 4D spatio-temporal space, i.e., x, y, z and the time series t, and then the action class is represented by the orientation of the normal in the 4D space. Such HON4D feature is more discriminative than previous approaches. However, it does not capture the geometrical structure of the human, which is also meaningful and important for recognizing the activity of the human. In addition, Yang and Tian [32] introduce the SNV, which is relevant to our approach. In this approach, the polynormal is proposed, which well captures the surface cue of the human action. However, this approach also suffers from the issue that the geometrical structure of the surface is not embedded into the model. Lin et al. [11] introduce the depth and skeleton associated approach for recognition. Our approach is totally different from Lin et al. [11]. Firstly, Lin et al. [11] aim to improve the recognition accuracy of the color sequence, while our approach is designed for the depth sequence. Secondly, in Lin et al. [11], an auxiliary, multi-modal database is utilized for improving the recognition accuracy. In contrast, our approach does not use any auxiliary information. In addition, Lin et al. [11] do not carefully exploit the relationship between the skeleton and the appearance, while our approach well captures the appearance cue of the surface and the geometrical structure of different surfaces.

3. Problem formulation

Given the observed action sequence $\mathbf{A} = \{A_i \mid i = 1, 2, \ldots, I\}$, the goal of action recognition is to infer the class label $l_i \in \{1, 2, \ldots, L\}$ for each action sequence $A_i \in \mathbf{A}$. In special cases, an action sequence may contain several human activities. Following the previous work [26], we assume that each sequence contains only one human activity in this paper. Hence, the objective of the human action recognition is to compute the class label l^* by maximizing the posterior as:

$$l^* = \arg\max p(l|A_i). \tag{1}$$

As mentioned, based on the joints, a human subject can be decomposed into several approximate rigid parts. For instance, the non-rigid left arm can be separated into the left wrist and elbow. Each segmented part is approximately rigid. Since the appearance cues of these parts are conditionally independent, the posteriori probability in Eq. (1) can be further factorized into the observation for

Please cite this article as: Y. Zhou, A. Ming, Human action recognition with skeleton induced discriminative approximate rigid part model, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.07.025