Signal Processing: *Image Communication* 33 (2015) 29–40

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Informative joints based human action recognition using skeleton contexts

Min Jiang [a,*], Jun Kong [a,b], George Bebis [c], Hongtao Huo [d]

[a] *Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China*
[b] *College of Electrical Engineering, Xinjiang University, Urumqi 830047, China*
[c] *Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557, United States*
[d] *Department of Information Security Engineering, People's Public Security University of China, Beijing 100038, China*

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form
7 February 2015
Accepted 10 February 2015
Available online 18 February 2015

Keywords:
Action recognition
Skeleton contexts
Informative joints
Affinity propagation
CRFs
Kinect

ABSTRACT

The launching of Microsoft Kinect with skeleton tracking technique opens up new potentials for skeleton based human action recognition. However, the 3D human skeletons, generated via skeleton tracking from the depth map sequences, are generally very noisy and unreliable. In this paper, we introduce a robust informative joints based human action recognition method. Inspired by the instinct of the human vision system, we analyze the mean contributions of human joints for each action class via differential entropy of the joint locations. There is significant difference between most of the actions, and the contribution ratio is highly in accordance with common sense. We present a novel approach named skeleton context to measure similarity between postures and exploit it for action recognition. The similarity is calculated by extracting the multi-scale pairwise position distribution for each informative joint. Then feature sets are evaluated in a bag-of-words scheme using a linear CRFs. We report experimental results and validate the method on two public action dataset. Experiments results have shown that the proposed approach is discriminative for similar human action recognition and well adapted to the intra-class variation.

1. Introduction

Automatic human action recognition has been a highly active research area in computer vision [1–5]. The main goal of human action recognition is to segment the human target from uncontrolled background and analyze the motion sequences to interpret the meaning of the action automatically. This research can be widely applied in various domains, such as public security surveillance, virtual reality, and computer games. So far, most of the research work mainly focuses on action recognition of intensity video sequence captured by RGB cameras. However, intensity images are sensitive to illumination. Observed from different viewpoints, the same action presents very different resulting images. Self-occlusion makes this problem even worse. Especially in the case of clutter background, segmentation of the human body is a very challenging task. Human action involves dynamic spatial and temporal information. Even if the human body is segmented accurately, the difficulty occurs due to the complexity of human actions. Moreover, human actions are closely influenced by different culture, personal character and emotion shift. How to reveal latent intra-class spatial-temporal law of each action with considerable intra-class variability and inter-class overlap becomes the primary issue for action recognition.

Depth map has drawn much interest for human action recognition [6–19]. Depth map records the distance from the surface of the object to the camera. With the depth information, human body can be detected and segmented robustly. In particular, depth based human skeleton tracking technology [6,20] achieves outstanding precision and

* Corresponding author.
E-mail address: minjiang@jiangnan.edu.cn (M. Jiang).

http://dx.doi.org/10.1016/j.image.2015.02.004
0923-5965/

stimulates the researches of human action recognition using skeleton information.

Our focus in this paper is to establish a robust scheme for human action recognition based on estimated skeletons only (Fig. 1). Specifically, we evaluate the contribution of all the joints and construct informative joint set for each action class to eliminate the disturbance of unrelated joints. To make our representation robust against variation of human body size and orientation, we retargeted the skeletons to a standard skeleton and normalize the skeletons by translation, rotation and scaling. Similarities between postures are evaluated by skeleton contexts, a binned pairwise spacial distribution of informative joints. To improve the robustness, we propose using multi-scale bins. We perform the quantization with AP [21] (Affinity Propagation) method to cluster the feature vectors into n (n is determined by preferences) posture vocabularies. Encoded sequential features are trained upon linear CRFs. Experiments show that the recognition performance achieves high precision on two public action databases: MSRAction3D [13] and UTKinect [16]. It is robust to intra-variations in viewpoints, performance styles and individual body sizes. It also has good quality to distinguish inter-similarity between different action classes.

This paper is organized as follows: Section 2 briefly reviews related work in action recognition over depth map. Section 3 discusses and analyzes the informative joints based feature extraction using skeleton contexts. In Section 4, we use linear CRFs to classify the action samples with the proposed representation. Section 5 presents experimental results and discussions of the proposed approaches. Finally, conclusions are drawn in Section 6.

2. Related work

With the development of depth sensors, especially the launching of Microsoft Kinect, there has been an upsurge

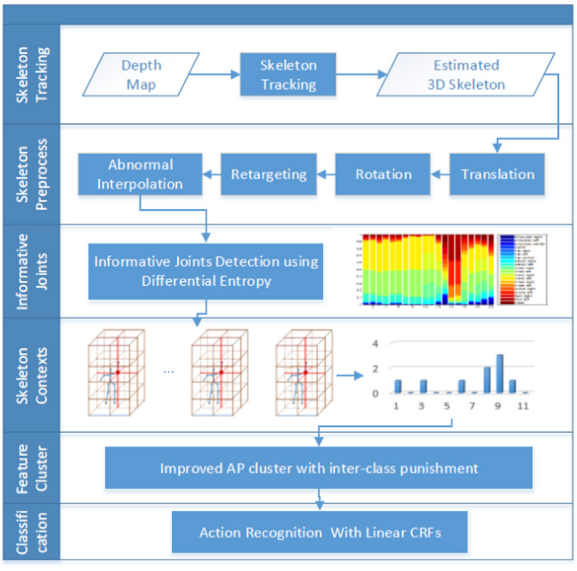

Fig. 1. Overview of the feature extraction and action classification scheme proposed in this paper.

of research on human recognition over depth map. Human action recognition using depth maps may be divided into two categories: algorithms using depth maps directly, and algorithms using estimated skeletons from depth maps.

In the first category, Lu et al. [11] introduce STIP's counterpart into depth video (called DSTIP). They describe the local 3D depth bin around the DSTIPs with a novel depth cuboid similarity feature (DCSF). DCSF features are clustered using K-means algorithm. Depth sequences are represented as a bag-of-codewords and classified by SVM with histogram intersection kernel. Oreifej et al. [12] describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. They trained the Histogram of Oriented 4D Normals (HON4D) using SVM with polynomial kernel. Li et al. [13] employ an action graph to model explicitly the dynamics of the actions and a bag of 3D points to characterize a set of salient postures that correspond to the nodes in the action graph. To reduce the computation cost, they project the depth map onto the three orthogonal Cartesian planes and further sample a specified number of points at equal distance along the contours of the projections. This sampling scheme is view dependant and leads to a poor accuracy of 74.7% in Cross Subject Test on MSRAction3D dataset. Dubey et al. [14] proposed to use depth camera and 3D-MHIs to model the a 3D human shape in space–time. The 3D-MHI approach combines the MHIs (Motion History Image) which encodes a range of times in a single frame, with two additional channels, fDMHIs (forward depth MHIs) and bDMHIs (backward depth MHIs). Experiments report a high precision of 97% upon RGB-D data comparing a lower precision of 87% upon traditional RGB data. Note that these experiments simply aim to detect falls from other actions. These algorithms are used to recognize activities without dependence on skeleton tracking.

In the second category, Ohn-Bar et al. [15] characterize actions using pairwise affinities between view-invariant joint angle features over the performance of an action. Using cosine distance function, this skeleton based method arrives at a precision of 83.53% on MSRAction3D dataset. Lu et al. [16] present a compact representation of postures with histograms of 3D joint locations (HOJ3D) and train the features by discrete hidden Markov models (HMMs). Ofli et al. [17] sort the joints by the highly interpretable measures such as the mean or variance of joint angle trajectories and automatically select a few most informative skeletal joints. The experiments demonstrate that the sequence of the most informative joints (SMIJ) reveals significant discrimination for most human actions. But it is insensitive to discriminate different planar motions around the same joint. This limitation leads to a low classification rate on MSRAction3D dataset. Evangelidis et al. [18] encode the relative position of joint quadruples. This short, view-invariant descriptor is then represented by Fisher vectors and trained with a Gaussian mixture model. Sung et al. [22] use a two-layered maximum entropy Markov model (MEMM) to classify combined features of skeletal features, skeletal HOG image features, and skeletal HOG depth features. Lin et al. [23] reported high recognition accuracy (precision/recall of 97.7/97.2). They