Information Sciences 444 (2018) 20–35

Contents lists available at ScienceDirect

Information Sciences

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

Hierarchical topic modeling with pose-transition feature for action recognition using 3D skeleton data☆

Thien Huynh-The[a], Cam-Hao Hua[a], Nguyen Anh Tu[a], Taeho Hur[a], Jaehun Bang[a], Dohyeong Kim[a], Muhammad Bilal Amin[a,d], Byeong Ho Kang[b], Hyonwoo Seung[c], Soo-Yong Shin[a,*], Eun-Soo Kim[e], Sungyoung Lee[a,*]

[a] *Department of Computer Science & Engineering, Kyung Hee University (Global Campus), 1732 Deokyoungdae-ro, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, South Korea*
[b] *School of Computing and Information System, University of Tasmania, Hobart, TAS 7005, Australia*
[c] *Department of Computer Science, Seoul Women's University, 621 Hwarang-ro, Gongneung 2(i)-dong, Nowon-gu, Seoul, South Korea*
[d] *National Research Foundation of Korea, 201 Gajeong-ro, Yuseong-gu, Daejeon 34113, South Korea*
[e] *Department of Electronic Engineering, Kwangwoon University, Seoul 01897, South Korea*

ARTICLE INFO

Article history:
Received 19 May 2017
Revised 14 February 2018
Accepted 18 February 2018
Available online 22 February 2018

Keywords:
3D action recognition
Topic modeling
Pose-transition feature
Pachinko allocation model
Depth camera

ABSTRACT

Despite impressive achievements in image processing and artificial intelligence in the past decade, understanding video-based action remains a challenge. However, the intensive development of 3D computer vision in recent years has brought more potential research opportunities in pose-based action detection and recognition. Thanks to the advantages of depth camera devices like the Microsoft Kinect sensor, we developed an effective approach to in-depth analysis of indoor actions using skeleton information, in which skeleton-based feature extraction and topic model-based learning are two major contributions. Geometric features, i.e. joint distance, joint angle, and joint-plane distance are calculated in the spatio-temporal dimension. These features are merged into two types, called pose and transition features, and then are provided to codebook construction to convert sparse features into visual words by *k*-means clustering. An efficient hierarchical model is developed to describe the full correlation of feature - poselet - action based on Pachinko Allocation Model. This model has the potential to uncover more hidden poselets, which have been recognized as the valuable information and help to differentiate pose-sharing actions. The experimental results on several well-known datasets, such as MSR Action 3D, MSR Daily Activity 3D, Florence 3D Action, UTKinect-Action 3D, and NTU RGB+D Action Recognition, demonstrate the high recognition accuracy of the proposed method. Our method outperforms state-of-the-art methods in the field in most dataset benchmarks.

☆ This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion). This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00655). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2011-0030079). This research was supported by Korea Research Fellowship program funded by the Ministry of Science, ICT and Future Planning through the National Research Foundation of Korea(NRF-2016H1D3A1938039).

* Corresponding authors.

E-mail addresses: thienht@oslab.khu.ac.kr (T. Huynh-The), hao.hua@oslab.khu.ac.kr (C.-H. Hua), tunguyen@khu.ac.kr (N. Anh Tu), hth@oslab.khu.ac.kr (T. Hur), jhb@oslab.khu.ac.kr (J. Bang), dhkim@oslab.khu.ac.kr (D. Kim), m.b.amin@ieee.org (M.B. Amin), byeong.kang@utas.edu.au (B.H. Kang), hwseung@swu.ac.kr (H. Seung), sooyong.shin@khu.ac.kr (S.-Y. Shin), eskim@kw.ac.kr (E.-S. Kim), sylee@oslab.khu.ac.kr (S. Lee).

https://doi.org/10.1016/j.ins.2018.02.042

1. Introduction

Human action analysis and understanding have been developed for human-machine interaction, video-based surveillance and monitoring, health and wellness assistance, sports coaching assistance, and robotic control systems [14]. Although many remarkable outcomes have been reported, viewpoint variation, occlusion, and multi-object interference [18] are still major challenges faced whenever proposing an efficient action recognition approach. The aspects of often considered for action recognition are input sensory data and feature-based action modeling due to their significant influence on the overall system performance in terms of recognition accuracy and processing speed. A depth camera with evident advantages is considered in this study to overcome the natural limitations of a traditional color camera for the purpose of pose assessment and action analysis [29]. Complementary information regarding depth and skeleton channels provided by the Kinect sensor [6] has brought about a reasonable solution for the considered channels. Combining the two different modal sensors of the depth camera and inertial body sensor was further studied to improve recognition accuracy.

Fundamentally, human actions consist of two main classes: the single action class and the interaction class. Single actions are usually performed by a person outdoors, such as walking, jogging, running, hand waving, jumping, and indoors, such as hand clapping, hand catching, and high warm waving. Some indoor human-object interactions, such as using a laptop, using a vacuum cleaner, making a phone call, and reading a book, were also investigated. Interactive actions, generally performed by two people in a scene, such as hand shaking, kicking, punching, pushing, pointing, and hugging, were prevalently discussed for both RGB and depth videos. The above approaches were mostly developed for assisted living in private areas and secure surveillance in public areas. Although they had the ability to detect and recognize simple actions under noise-free conditions, the recognition accuracy for complex actions in such complicated environments as multiple-view and occlusion should be improved appreciably. Most approaches lack a powerful learning model that is capable of precisely and robustly characterizing human actions using skeleton-based spatio-temporal features precisely and robustly. Therefore, the combination of skeleton-based action recognition and topic learning models is clearly an open research area.

Commercialization of the low-cost Microsoft Kinect sensor, first developed for console gaming, has been applied to more industrial and research opportunities. The advanced specifications of depth and visual information provided by Kinect potentially address many of the remaining problems of video-based action recognition. Currently, Kinect is researched and developed for widespread applications in which human-computer interactions are strongly emphasized. The fact that thousands of scientific publications and technical demonstrations have been delivered during the more than four years since the released date proves the huge benefits of the Kinect sensor.

Based on the many noted benefits of the Kinect sensor, in this paper, we develop an efficient method, that exploits 3D skeleton data for recognizing human actions. First, we extract the joint distance, joint angle, and joint-plane distance as geometric features by joint location on the 3D coordinate axis in the currently considered frame and the previous frame. These features are capable of describing the relationships between interactive body parts in space; however, the body motions over time are missed, leading to the incapability of deeply understanding the whole action. The fact that two or more different actions can share similar postures cause misclassification. For example, *write on paper* and *use laptop* are two different actions; however, their appearance is mostly characterized by the *sitting* posture. To handle such issues, a flexible hierarchical topic model developed on the Pachinko Allocation Model (PAM) is capable of representing the relationship of feature-poselet-action in multi-frame observation [15]. Geometric features are fused into the pose and transition features before being clustered and mapped to visual codewords. Unlike the existing approaches, which map individual features, neither distance nor angle, to codewords, our method encodes merged features to strengthen the posture differentiation. Utilizing Directed Acyclic Graph (DAG), PAM potentially graphs not only the correlations between features but also those of poselets and actions to enrich the action categorization. Finally, the actions are recognized by an advanced Support Vector Machine (SVM) classifier using a χ^2 kernel. The proposed method is evaluated and compared to modern approaches using five well-known datasets of 3D action recognition: MSR Action 3D, MSR Daily Activity 3D, Florence 3D Action, UTKinect-Action 3D, and NTU RGB+D Action Recognition.

The remainder of this paper is arranged as follows. Existing RGB- and depth-video action recognition approaches are briefly reviewed in Section 2. Section 3 introduces the proposed action recognition. The experimental evaluations and results are discussed in Section 4. Finally, the conclusion and research orientation for future work are given in Section 5.

2. Related work

As remarkable benefits, the Kinect sensor is able to provide depth, RGB, and body map channels simultaneously. Therefore, visual features extracted from depth cameras can be categorized into two major classes. The first class includes skeleton-based features extracted from 3D coordinates of body parts for action recognition. The most popular and simple feature is the 3D skeleton trajectory [1,4,13,26,35,46] to temporally explain body transitions. To determine body movement and rotation challenges, Seidenari et al. [26] designed a volumetric-temporal feature descriptor to extract the overall discriminative features from the depth map dataset. Each person in [13] was described by a set of 3D edge vectors connecting joints within a frame and another set of 3D trajectory vectors connecting joints of several previous frames. Based on recognizing 3D body motions as elements of an exceptional Euclidean group $SE(3)$, Vemulapalli et al. [35] encoded actions