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
0020-0255/© 2018 Elsevier Inc. All rights reserved.

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 to curves in a Lie group to classify using one-vs-all linear SVM. In addition to the skeleton trajectory, 3D efficient feature descriptors were also developed for action recognition. Devanne et al. [4] addressed a problem of joint trajectory similarity calculation in a Riemannian manifold by warping trained 3D points to query trajectories and extracting elastic metrics between trajectory shapes. The skeleton trajectory was extensively studied for action presentation based on Kendall's shape manifold [1], which aims to solve data corruption in many execution rates within and across subjects during the collection progress. A parameterization-invariant metric formulated by combining a standard Euclidean norm and transported square-root vector fields retrieved from trajectories is evaluated for the transformation of trajectories to obtain computational efficiency. Zhao et al. [49] developed the structured streaming skeleton (SSS), an efficient feature extraction that facilities modeling of a skeleton within a frame to a vector containing major attributes. EigenJoint [42], depicted by principal component analysis (PCA) on 3D skeleton data to construct a static pose, a gesture feature, and he entire dynamics, was offered for action modeling. Chen and Guo [3] proposed TriViews, which allows for the projection of 3D depth maps to the following three viewpoints: front, side, and top. The combination of the three views produces a full description of an action for recognition. To reduce latency in action recognition systems, Cai et al. [2] advanced a skeleton-based action representation that was capable of aggregating encoded features of individual limbs using a Markov random field. Recently, Shahroudy et al. [28] proposed a joint sparse regression-based learning approach to encode each action as a combination of multi-modal features which were constructed from a sparse set of body parts. These aforementioned approaches were able to guarantee cheap complexity; however, recognition accuracy was obviously influenced by 3D skeleton data where the joint localization accuracy is decided by estimation techniques.

Another group consists of visual features that are directly extracted from depth sequences. In [25], Oreifej and Liu captured surface normal directions, which are distributed in the spatial, depth, and temporal dimension by a histogram. For descriptiveness and discrimination increments, body surface context (BSC) [31] was introduced to express the locally relative distribution of neighbors corresponding to a reference joint. Moreover, the skeleton-based, random-reference-point, and spatio-temporal schemes were developed based on BSC features for action representation. To deal with variations in intra-class gestures by various manners and length of actions, hierarchical 3D kernel descriptors studied by Kong and Fu [12] are efficient to calculate raw pixel-base traits, low-level patch-base features, and mid-level video-based features. Shape and motion features [10], extracted from object silhouette information in the depth images, were provided to hidden Markov models (HMMs) for model training and recognition. Since only the depth channel is requested for visual feature extraction, most approaches cannot report a competitive result in term of accuracy.

Regarding the less valuable information obtained from only the depth channel, RGB-D hybrid features have been recently considered to raise recognition accuracy. Ohn-Bar et al. [24] upgraded a histogram of oriented gradients (HOG) to a dual-layer feature descriptor for the task of feature extraction in the spatio-temporal domain. Liu et al. [19] fused RGB and depth material by conditional random fields under coupled hidden context. The collaboration between RGB and depth modalities is exploited over a distinctive graph structure designed by potential functions. From RGB and depth videos, Song et al. [30] studied trajectories of surface patches (ToSP) to depict the shape and position variations of human bodies. To capture the complex correlations containing shared features from multi-modal data of color and depth channels, Kong and Fu [11] proposed a discriminative relational representation learning (DRRL) approach. A hinge lost function was furthermore introduced to enrich the discrimination of relational features by measuring classification loss. Recently, an efficient 4D color-depth (CoDe4D) local spatio-temporal (LST) feature characterizing texture, shape, and pose variations was introduced by Zhang et al. [47] to incorporate intensity and depth information. High accuracy is the essential goal of RGB-D-based approaches, but processing latency and big data storage have become practical challenges.

Several researchers have approached the recognition problem via action modeling techniques. Luo et al. [22] developed a sparse coding-based temporal pyramid matching approach (ScTPM) and a robust center-symmetric motion local ternary pattern (CS-Mltp) descriptor to extract features of 3D-joints as well as CS-Mltp from depth maps and RGB videos, respectively. A hidden layer was designed by Su et al. [32] for the latent attribute model to control the concurrence of part-wise and body-wise attributes. Moreover, the overall system latency is potentially reduced by aggregating modeled attributes of individual skeleton limbs. In order to suppress noise and overcome incorrect joint tracking issues with a depth camera, Wang et al. [37] studied an efficient actionlet ensemble model that proficiently captured the correlations between subsets of human joints. Moreover, this model was invariant to transitional and temporal misalignments when a person continuously moves or interacts with an object. In [23], activities were clustered using a *k*-means-based unsupervised approach and recognized by HMMs. Liu et al. [20] formulated the task of action recognition in single- and multiple-view conditions to a part-regularized multitask structural learning model. The approach has an ability to encounter feature subspaces of specific- and shared-oriented actions to enhance the overall recognition accuracy through the extension of learning capacity. Du et al. [5] designed a novel hierarchical recurrent neural network (RNN) that is capable of working with fused hierarchical local features. Shahroudy et al. [27] and Liu et al. [21] introduced long short-term memory (LSTM) networks for modeling the correlations of the features for each body part using 3D skeleton data. An impressive approach, proposed by Wang et al. [38], combined weighted hierarchical depth motion maps (WHDMM) and a three-channel deep convolutional neural network (3ConvNets), in which 2D spatial structures, encoded from the spatio-temporal motion patterns by WHDMMs at different temporal scales, were independently supplied to 3ConvNets with ImageNet models. A drawback of most of the investigated approaches is the unavailable or poor description of the correlation between features, postures, and actions in space and time.

The topic model has been discussed in several recent works [7–9,33,40]. Wang et al. [40] developed an Image-Dominant Topic Model (IDTM) for text-semantic mapping and incorporated the IDTM with Convolutional Neural Network (CNN) for text-image relationship establishment. For the large-scale image retrieval applications, Tu et al. [33] formalized the spatial structure among the local features by a probabilistic topic model that efficiently allows extracting the object and background regions from an image. Huynh-The et al. presented the methods for interactive action recognition, in which they extract only the joint distance metric [7] and then complement the joint angle metric [9] from the skeleton coordinate information in the spatio-temporal dimension. Due to utilizing one codebook to encode two separated types of codewords, PAM is unable to capture the internal correlation of each codeword set, which contains a pair of joint distance and joint angel metrics. Therefore, the fact that the induced discrimination among poselets in the upper layer is insignificant leads to low recognition accuracy in the case of complex actions. Different from [7,9], a merged feature of joint distance and angle metrics [8] is encoded as a single codeword. In other words, the PAM-based topic model is built to fit only one kind of codeword, which can potentially deliver high accuracy thanks to the highly distinctive merged feature consisting of the spatio-temporal information of the joint distance and angle. In this work, the topic model is built to flexibly fit two new types of codewords, i.e., pose and transition. These newly defined codeword pairs are clustered by two independent codebooks. Concretely, the pose codeword contains posture information, including the joint distance metric, joint angle metric, and joint-plane distance metric of a complete skeleton in a frame, while the transition codeword contains the movement information of a skeleton in time. Since these two types of codewords are modeled independently and simultaneously, the model has the ability to entirely comprehend the correlation not only inside each codeword pair of pose and transition but also among distinctive pairs of pose-transition. Consequently, the distinctiveness margin among action classes is enlarged for improving recognition accuracy and some hidden peculiar actions are further discovered.

3. Methodology

The proposed action recognition approach has five modules as follows: 3D skeleton acquisition from the Kinect sensor, spatio-temporal feature extraction, codebook construction, topic modeling, and SVM classification.

3.1. Skeleton acquisition

In this work, 3D skeleton data containing the body joint coordinates of detected people in a scene is provided by Kinect sensor. Besides the RGB and depth channels, Kinect sensor also supports the body frame channel with moderate accuracy of pose estimation. Most of existing researches have studied on the datasets captured by Kinect v1, released in 2012, such as MSR Action 3D [16], MSR Daily Activity 3D [36], Florence 3D Action [26], UTKinect-Action 3D [41], and SBU Kinect Interaction [45]. In 2014, Microsoft released the update version for Kinect sensor, called Kinect v2, with the valuable upgrade in hardware and software specifications, especially the pose estimation algorithm. For instance, the resolutions of the color and depth frames are improved from 640 × 480 to 1920 × 1080 and from 320 × 240 to 512 × 424, respectively. Regarding pose estimation, the number of skeleton joints is upgraded from 20 to 25 with more accurate tracking and six full skeletons in maximum are tracked simultaneously. However there is only one 3D action recognition dataset generated by Kinect v2, NTU Action Recognition from Nanyang Technological University [27], which has been published for free access. High system requirement is now a practical challenge for researchers and developers who are interested in Kinect v2.

3.2. Skeleton-based spatio-temporal feature extraction

In this stage, skeleton-based features are extracted to describe human posture in the spatial dimension and to capture pose movement in the temporal dimension. Concretely, we calculate three common metrics mostly used for action recognition: the pair-wise joint distance, the joint-joint vectors versus three coordinate axes angle, and the joint-plane distance where the plane here is constructed by three joints other than the mentioned one. Each joint belonging to a full skeleton at tth frame is given as $p^t = \left(x^t, y^t, z^t\right)$ in space $\Re^3$.

Joint distance: is determined as the 3D Euclidean distance of an arbitrary pair of joints i and j. As mentioned before, we calculate this metric for such joints within a frame, denoted d_{ji}^t as pose feature representation, and for such joints in two consecutive frames, denoted $d_{ji}^{\Delta t}$ as transition feature representation

$$\begin{aligned} d_{ji}^t &= \left\| p_j^t - p_i^t \right\| \\ &= \sqrt{\left(x_j^t - x_i^t\right)^2 + \left(y_j^t - y_i^t\right)^2 + \left(z_j^t - z_i^t\right)^2} \\ d_{ji}^{\Delta t} &= \left\| p_j^{t-1} - p_i^t \right\| \\ &= \sqrt{\left(x_j^{t-1} - x_i^t\right)^2 + \left(y_j^{t-1} - y_i^t\right)^2 + \left(z_j^{t-1} - z_i^t\right)^2} \end{aligned} \tag{1}$$

With a full detected skeleton consisting of m joints, there are $m(m-1)/2$ values of d_{ji}^t $(i \neq j)$, and m^2 values of $d_{ji}^{\Delta t}$.

Joint angle: is defined as the angle between a joint-joint vector $\overrightarrow{ji}$ versus unit vectors of the horizontal axis $\overrightarrow{Ox} = \langle 1, 0, 0 \rangle$ (denoted θ_{jix}), the vertical axis $\overrightarrow{Oy} = \langle 0, 1, 0 \rangle$ (denoted θ_{jiy}), and the depth axis $\overrightarrow{Oz} = \langle 0, 0, 1 \rangle$ (denoted θ_{jiz}). Similar to joint distance, we also extract pose feature $\theta^t_{ji} = (\theta^t_{jix}, \theta^t_{jiy}, \theta^t_{jix})$ with $\overrightarrow{ji^t} = \langle x^t_i - x^t_j, y^t_i - y^t_j, z^t_i - z^t_j \rangle$

$$\begin{aligned}
\theta^t_{jix} &= \angle\left(\overrightarrow{ji^t}, \overrightarrow{Ox}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^t} \cdot \overrightarrow{Ox}}{\left\|\overrightarrow{ji^t}\right\| \left\|\overrightarrow{Ox}\right\|}\right) \\
\theta^t_{jiy} &= \angle\left(\overrightarrow{ji^t}, \overrightarrow{Oy}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^t} \cdot \overrightarrow{Oy}}{\left\|\overrightarrow{ji^t}\right\| \left\|\overrightarrow{Oy}\right\|}\right) \\
\theta^t_{jiz} &= \angle\left(\overrightarrow{ji^t}, \overrightarrow{Oz}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^t} \cdot \overrightarrow{Oz}}{\|\overrightarrow{ji^t}\| \|\overrightarrow{Oz}\|}\right)
\end{aligned} \tag{2}$$

and transition feature $\theta^{\Delta t}_{ji} = (\theta^{\Delta t}_{jiz}, \theta^{\Delta t}_{jiy}, \theta^{\Delta t}_{jiz})$ with $\overrightarrow{ji^{\Delta t}} = \langle x^t_i - x^{t-1}_j, y^t_i - y^{t-1}_j, z^t_i - z^{t-1}_j \rangle$

$$\begin{aligned}
\theta^{\Delta t}_{jix} &= \angle\left(\overrightarrow{ji^{\Delta t}}, \overrightarrow{Ox}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^{\Delta t}} \cdot \overrightarrow{Ox}}{\left\|\overrightarrow{ji^{\Delta t}}\right\| \left\|\overrightarrow{Ox}\right\|}\right) \\
\theta^{\Delta t}_{jiy} &= \angle\left(\overrightarrow{ji^{\Delta t}}, \overrightarrow{Oy}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^{\Delta t}} \cdot \overrightarrow{Oy}}{\left\|\overrightarrow{ji^{\Delta t}}\right\| \left\|\overrightarrow{Oy}\right\|}\right) \\
\theta^{\Delta t}_{jiz} &= \angle\left(\overrightarrow{ji^{\Delta t}}, \overrightarrow{Oz}\right) = \cos^{-1}\left(\frac{\overrightarrow{ji^{\Delta t}} \cdot \overrightarrow{Oz}}{\left\|\overrightarrow{ji^{\Delta t}}\right\| \left\|\overrightarrow{Oz}\right\|}\right)
\end{aligned} \tag{3}$$

Joint-plane distance: is termed as the distance from a joint i to a plane formed by three joints j, k, and l. At first, the normal vector to plane f_{jkl} is given

$$\begin{aligned}
\overrightarrow{n_{jkl}} &= \overrightarrow{u} \times \overrightarrow{v} \\
&= \left\langle \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \right\rangle \\
&= \langle a, b, c \rangle
\end{aligned} \tag{4}$$

where $\overrightarrow{u} = \overrightarrow{jk}$ and $\overrightarrow{v} = \overrightarrow{jl}$ are defined as

$$\begin{aligned}
\overrightarrow{u} &= \langle x_k - x_j, y_k - y_j, z_k - z_j \rangle \\
&= \langle u_1, u_2, u_3 \rangle \\
\overrightarrow{v} &= \langle x_l - x_j, y_l - y_j, z_l - z_j \rangle \\
&= \langle v_1, v_2, v_3 \rangle
\end{aligned} \tag{5}$$

The scalar equation of plane f is outlined as $ax + by + cz + d = 0$. Then vector from point i to plane f is addressed as:

$$\vec{w} = -\langle x - x_i, y - y_i, z - z_i \rangle \tag{6}$$

The joint-plane distance is calculated for the case of pose feature

$$\partial^t_{ijkl} = \frac{\left|n^t_{jkl} \cdot w\right|}{\left\|n^t_{jkl}\right\|} = \frac{\left|ax^t_i + by^t_i + cz^t_i + d\right|}{\sqrt{a^2 + b^2 + c^2}} \tag{7}$$

and for the case of transition feature where joint i belongs to tth frame and three joints j, k, l for plane construction belong to $(t-1)^{th}$ frame

$$\partial^{\Delta t}_{ijkl} = \frac{\left|n^{t-1}_{jkl} \cdot w\right|}{\left\|n^{t-1}_{jkl}\right\|} \tag{8}$$

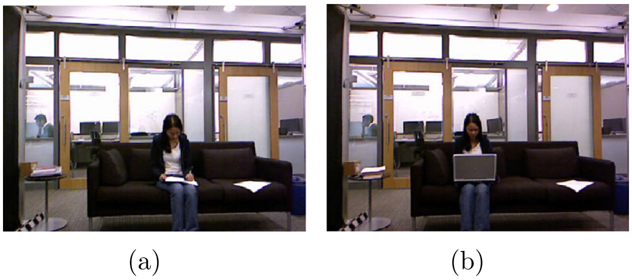


Fig. 1. Example of two distinctive actions having the same *sitting* posture: (a) Write on paper, (b) Use laptop.

Given an arbitrary joint, two merged features, pose feature denoted c_Δ and transition feature denoted c_∇, consisting of normalized distance, angle, and joint-plane distance metrics are organized as follows

$$\begin{aligned} c_\Delta &= \left\{ \widehat{d^t}, \widehat{\theta^t}, \widehat{\partial^t} \right\} \\ c_\nabla &= \left\{ \widehat{d^{\Delta t}}, \widehat{\theta^{\Delta t}}, \widehat{\partial^{\Delta t}} \right\} \end{aligned} \tag{9}$$

where the normalized metrics are generally defined as

$$\begin{aligned} \widehat{d} &= \frac{d - \min(d)}{\max(d) - \min(d)} \\ \widehat{\theta} &= \frac{\theta}{2\pi} \\ \widehat{\partial} &= \frac{\partial - \min(\partial)}{\max(\partial) - \min(\partial)} \end{aligned} \tag{10}$$

At this point, m pose feature vectors c_Δ and m transition feature vectors c_∇ are totally extracted for a captured skeleton at each frame. It is noted that the dimensions of c_Δ and c_∇ are not equal due to different number of joint distance values.

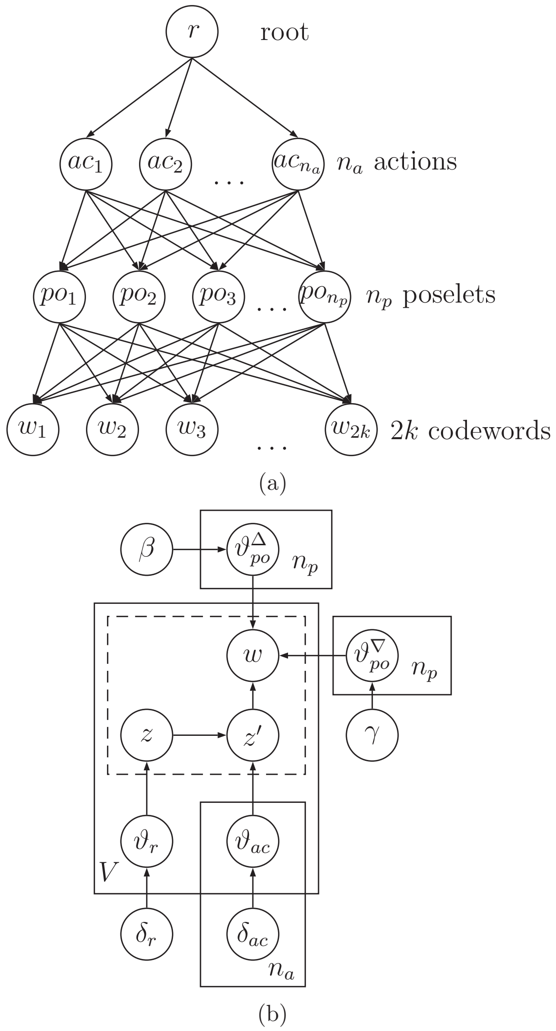
3.3. Codebook construction

Encoding skeleton-based features to visual codewords is a preprocessing step of several topic modeling techniques. In this research, for visual codebook construction, we exploit k-means clustering technique which utilizes the metric of Euclidean distance for partitioning features into k clusters. As mentioned before, the dimensions of pose feature and transition feature vectors are different; therefore we cluster them separately, i.e., there are two codebooks built. To be more specific, two types of codeword corresponding to c_Δ and c_∇, called pose codeword and transition codeword, are produced. The number of clusters, a.k.a. the codebook size, is given in advance, however, appropriate selection of this parameter is sometimes ambiguous so that an acceptable trade-off between classification error and computational cost is approached. Additionally, the risk of over-fitting issue may occurs unexpectedly.

3.4. Topic modeling

The merged features containing the information of pose state and movement used for action presentation are converted to codewords. Instead of simply pushing such the codeword histogram to common classifiers to recognize actions in a short duration, a longstanding observation is studied and analyzed for complicated activities, such as *write on paper* with the *sitting* pose as shown in Fig. 1. Although the topic model techniques are fundamentally proposed to solve many high-challenging tasks relating to the natural language processing field, they can also be exploited for such image processing and computer vision issues as semantic-based image retrieval [33], cross-media topic detection and analysis [40], image understanding [17], and interactive activity recognition [9]. In this work, we construct a hierarchical model by adapting the idea of Pachinko Allocation Model, which has an ability to ascertain the feature-poselet-action correlation. Originally, PAM depicts one distribution mixture for only one corresponding single set of topics in which topic co-occurrences are described entirely. Each interior node inside the model graph is presented by a Dirichlet distribution over lower level nodes. The simplest version contains only one single layer of Dirichlet distributions between the root at the top level with many codeword distributions at the bottom level considerately.

PAM is able to efficiently expose and learn arbitrary, hidden, and sparse correlations of complicated relations between codewords and mined topics thanks to its essential specifications inherited from Direct Acyclic Graphs. Although PAM is firstly promoted with an arbitrary number of middle levels which have responsibility for capturing top-bottom relations, the four-level structure presented a pleasant trade-off between structural complexity and modeling efficiency [15]. PAM-based hierarchical topic model developed for action recognition is progressed with the top level representing a root action r, the second level AC characterizing n_a actions such that $AC = \{ac_1, ac_2, \ldots, ac_{n_a}\}$, the third level PO characterizing n_p poselets



(a)

(b)

Fig. 2. The proposed PAM-based hierarchical topic model used for action modeling: (a) 4-level model structure (b) graphical illustration. For each video, the model depicts multinomials ϑ_r and ϑ_{ac} from Dirichlet distributions at the root and the actions. For each codeword w, our model depicts an action z from ϑ_r, a poselet z' from ϑ_{ac} and then the codeword, including the pose codeword from multinomial distribution ϑ^{Δ} and the transition codeword from multinomial distribution ϑ^{∇}, is associated with z'. Compared with original PAM, the proposed topic model that is developed to flexibly and simultaneously learn two types of codewords is capable of portraying actions more discriminatively.

such that $PO = \{po_1, po_2, \ldots, po_{n_p}\}$, and the bottom level CW signifying $2k$ unique pose and transition codewords such that $CW = \{cw_1, cw_2, \ldots, cw_{2k}\}$. According to the 4-level hierarchy structure as shown in Fig. 2(a), the associations are directly built between the root and n_a actions, between n_a actions and n_p poselets, and between n_p poselets and $2k$ codewords. With a single Dirichlet distribution $Dir_r(\delta_r)$ assigned to the root r, we firstly depict the multinomial distribution $\vartheta_r^{(t)}$, where t is the considered body frame and δ_r is the parameter of the Dirichlet prior on per-image action distributions. Similarly, we also write the multinomial distributions $\vartheta^{(t)}_{ac_i|_{i=1}^{n_a}}$ over poselets from Dirichlet distributions $Dir_{ac}(\delta_{ac_i})|_{i=1}^{n_a}$ assigned to n_a actions. The poselets are encoded by $\vartheta^{\Delta}_{po_i|_{i=1}^{n_p}}$ and $\vartheta^{\nabla}_{po_i|_{i=1}^{n_p}}$ which are fixed multinomial distributions correspondingly depicted from Dirichlet distributions $Dir(\beta)$ and $Dir(\gamma)$ of pose and transition codewords for the whole video, respectively, where β and γ are the parameters of Dirichlet prior on per-poselet codeword distributions. For each codeword w captured in the body frame t, we finally depict an action z_w from $\vartheta_r^{(t)}$, a poselet z'_w from $\vartheta^{(t)}_{z_w}$, and codeword w from $\vartheta^{\Delta}_{z'_w}$ and $\vartheta^{\nabla}_{z'_w}$. The graphical illustration of our proposed hierarchical topic model for processing with two types of codeword is drawn in Fig. 2(b).

According to the above processing steps, the joint probability of the generation of a frame t, the action assignments $z^{(t)}$, the poselet assignments $z'^{(t)}$, and the multinomial distribution $\vartheta^{(t)}$ conditioned on Dirichlet distributions is expressed as follows

$$P\Big(t, z^{(t)}, z'^{(t)}, \vartheta^{(t)}\Big|\delta, \beta, \gamma\Big) = P\Big(\vartheta_r^{(t)}\Big|\delta_r\Big) \times \prod_{i=1}^{n_a} P\Big(\vartheta_{ac_i}^{(t)}\Big|\delta_{ac_i}\Big)$$
$$\times \prod_w \Big\{P\Big(z_w|\vartheta_r^{(t)}\Big)P\Big(z'_w\Big|\vartheta_{z_w}^{(t)}\Big)P\Big(w|\vartheta_{z'_w}^{\Delta}, \vartheta_{z'_w}^{\nabla}\Big)\Big\} \tag{11}$$

where $P(w|\vartheta_{po_w}^{\Delta}, \vartheta_{po_w}^{\nabla}) = P(w|\vartheta_{po_w}^{\Delta})P(w|\vartheta_{po_w}^{\nabla})$. The likelihood of body frame t is delivered by integrating out $\vartheta^{(t)}$ and summing over $ac^{(t)}$ and $po^{(t)}$ as follows

$$P(t|\delta, \beta, \gamma) = \int P\Big(\vartheta_r^{(t)}\Big|\delta_r\Big) \prod_{i=1}^{n_a} P\Big(\vartheta_{ac_i}^{(t)}\Big|\delta_{ac_i}\Big)$$
$$\prod_w \sum_{z_w, z'_w} \Big\{P\Big(z_w|\vartheta_r^{(t)}\Big)P\Big(z'_w\Big|\vartheta_{z_w}^{(t)}\Big)P\Big(w|\vartheta_{z'_w}^{\Delta}, \vartheta_{z'_w}^{\nabla}\Big)\Big\}d\vartheta^{(t)} \tag{12}$$

According to the reflection of several frames, we write the probability of generating a video $V = \{t_1, t_2 \ldots, t_N\}$, where N is the number of frames in V as follows

$$P(V|\delta, \beta, \gamma) = \int \prod_{i=1}^{n_p} \Big\{P\Big(\vartheta_{po_i}^{\Delta}|\beta\Big)P\Big(\vartheta_{po_i}^{\nabla}|\gamma\Big)\Big\} \prod_t P(t|\delta, \beta, \gamma) \tag{13}$$

The joint probability $P(V, AC, PO|\delta, \beta, \gamma.)$ of the video V and the action and poselet assignments is formulated as

$$P(V, \mathbf{z}, \mathbf{z}'|\delta, \beta, \gamma) = P(\mathbf{z}|\delta) \times P(\mathbf{z}'|\mathbf{z}, \delta) \times P(V|\mathbf{z}', \beta) \times P(V|\mathbf{z}', \gamma) \tag{14}$$

where the above terms are defined by integrating out the sampled multinomials

$$P(\mathbf{z}|\delta) = \int \prod_t P\Big(\vartheta_r^{(t)}|\delta_r\Big) \prod_w P\Big(z_w\Big|\vartheta_r^{(t)}\Big)d\vartheta$$
$$P(\mathbf{z}'|z, \delta) = \int \prod_t \left(\prod_{i=1}^{n_a} P\Big(\vartheta_{ac_i}^{(t)}|\delta_{ac_i}\Big) \prod_w P\Big(z'_w\Big|\vartheta_{z_w}^{(t)}\Big)\right)d\vartheta$$
$$P(V|\mathbf{z}', \beta) = \int \prod_{i=1}^{n_p} P\Big(\vartheta_{po_i}^{\Delta}|\beta\Big) \prod_t \left(\prod_w P\Big(w\Big|\vartheta_{z'_w}^{\Delta}\Big)\right)d\vartheta$$
$$P(V|\mathbf{z}', \gamma) = \int \prod_{i=1}^{n_p} P\Big(\vartheta_{po_i}^{\nabla}|\gamma\Big) \prod_t \left(\prod_w P\Big(w\Big|\vartheta_{z'_w}^{\nabla}\Big)\right)d\vartheta \tag{15}$$

It should be noted that we have to sample the subtopic assignments for each pose $P(V|\mathbf{z}', \beta.)$ and transition codeword $P(V|\mathbf{z}', \gamma.)$. The conditional distribution for action and poselet assignments is delivered as follows (16),

$$P\Big(z_w = ac_i, z'_w = po_j\Big|V, z_{-w}, z'_{-w}, \delta, \beta, \gamma\Big) \propto P\Big(w, z_w, z'_w\Big|V_{-w}, z_{-w}, z'_{-w}, \delta, \beta, \gamma\Big)$$
$$= \frac{P\Big(V, z, z'|\delta, \beta, \gamma\Big)}{P(V, z_{-w}, z'_{-w}|\delta, \beta, \gamma)}$$
$$= \frac{n_i^{(t)} + \delta_{ri}}{n_r^{(t)} + \sum_{ia=1}^{n_a} \delta_{ri}} \times \frac{n_{ij}^{(t)} + \delta_{ij}}{n_i^{(t)} + \sum_{j=1}^{n_p} \delta_{ij}}$$
$$\times \frac{n_{jk} + \beta_k}{n_j^{(t)} + \sum_{k=1}^{K} \beta_k} \times \frac{n_{jl} + \gamma_l}{n_j^{(t)} + \sum_{l=1}^{K} \gamma_l} \tag{16}$$

where $n_r^{(t)}$ is the number of occurrences of the root r in frame t; $n_i^{(t)}$ is the number of occurrences of action ac_i in frame t; $n_j^{(t)}$ is the number of occurrences of poselet po_j in frame t; $n_{ij}^{(t)}$ is the number of times that poselet po_j is sampled from action ac_i, and n_{jk} is the number of occurrences of pose codeword w_k^{Δ} in poselet po_j, and n_{jl} are the numbers of occurrences of transition codeword w_l^{∇} in poselet po_j. The notation $-w$ indicates all action assignments except codeword w, hence the numbers of occurrences do not cover w and their assignments. The hyper-parameters δ and β are estimated via the Gibbs sampling algorithm which is formulated in details as in [15]. The new data tagged by the pose and transition features, known as codewords, is produced as the output of PAM.

4. Experiment results and discussion

This section benchmarks our developed approach on five well-known 3D action recognition datasets which are collected by the Kinect sensor: MSR Action 3D [16], MSR Daily Activity 3D [36], Florence 3D Action [26], UTKinect-Action 3D [41], and NTU RGB+D Action Recognition [27]. The sensitivity is thoughtfully investigated with various parameter settings and further compared to various impressive methods in the field.

Table 1
Method sensitivity evaluation on the feature type impact.

Dataset	Feature type	
	Pose ($c_{\triangle}$)	Transition ($c_{\triangledown}$)
MSR Action 3D	90.48	93.04
MSR Daily Activity 3D	86.25	89.38
UTKinect-Action 3D	94.00	96.00
Florence 3D Action	90.70	92.09
NTU RGB+D (Cross-View)	72.58	76.28
NTU RGB+D (Cross-Subject)	65.95	68.17
Average	83.33	85.83

4.1. Dataset

MSR Action 3D: 557 videos presenting 20 different actions in which each subject is requested to play 2–3 times/action. Some sport-oriented actions contain a wide varying of body movements. Additionally, unstable motion speed in action execution should be addressed. Evaluating the dataset follows the cross-subject testing procedure [16], where the samples of odd-numbered subjects are applied for training and the rest for performance testing.

MSR Daily Activity 3D: Presents 16 indoor activities including single actions and human-object interactions that are classified into 10 subjects in which each subject contains actions in standing and sitting poses. Performing in widely spatial and temporal dimension make it a difficult challenge to precisely recognize the activities. Furthermore, the 3D joint coordinates estimated by the tracker are very noisy due to unguaranteed minimum depth. The cross-subject protocol in [36] is followed to benchmark this dataset.

UTKinect-Action 3D: Includes 200 videos of 10 actions which are recorded by 10 actors. Some body components are not located correctly in human-object interactions due to the occlusion that occurs the in case of hidden tracking. As with [41], we implement leave-one-out cross-validation for performance evaluation.

Florence 3D Action: Gathered at the University of Florence by the Kinect sensor and presents 9 activities in 215 sequences. In this dataset, the variation of object direction and motion velocity in human-object interactions needs to be considered to recognize them accurately. Using the same protocol in [26], the dataset is benchmarked by leave-one-subject-out cross-validation.

NTU RGB+D Action Recognition: Newly recorded by Kinect v2 with remarkable improvements in the hardware and software and consists of 56,880 samples covering 60 action classes including daily actions, health-related actions, and mutual actions. The variation of camera configurations in space aims to challenge the accuracy of action recognition. Following [27], cross-subject and cross-view protocols are appointed to assess our approach on this dataset.

4.2. Experimental setup

For each n-joint skeleton in a frame, we extract n pose feature vectors $c_{\triangle}$ and n transition feature vectors $c_{\triangledown}$. They are correspondingly encoded to pose and transition codewords by two codebooks having the same cluster amount, i.e., $k = 500$. The number of poselets of our hierarchical model is configured with $n_p = 200$. We set the parameter of the root of the fixed Dirichlet distribution at 0.01. The parameter of multinomial distributions over actions and poselets sampled from the Dirichlet distribution is fixed at 0.01. The N-class pattern recognition problem is handled by the SVM classifier using a χ^2 kernel [34]. The number of burn-in iterations for the Gibbs sampling is set at 1000. Additionally, for every 250 iterations, 50 samples are produced. Two experiments, simulated using MATLAB 2014b on a notebook computer with a 2.70-GHz Intel Core i7 CPU and 8GB RAM, are described as follows:

- The first experiment evaluates our approach under various parameter configurations to analyze its sensitivity.
- The second experiment aims to compare our proposed hierarchical topic model with state-of-the-art approaches in terms of recognition accuracy.

4.3. Results and discussion

4.3.1. Sensitivity analysis

The sensitivity of the proposed method is studied under various parameter configurations in which three parameters are considered: the feature type, the codebook size, and the number of poselets.

Feature type: Two feature types, $c_{\triangle}$, representing the posture information, and $c_{\triangledown}$, describing the transition information, correspond to two types of the codewords and are used separately in the method. They are constructed from one distance and three angle values for each pair of joints. Based on the quantitative results of recognition accuracy reported in Table 1, the transition feature is obviously better than the pose feature in all dataset tests, with an average accuracy that is 2.5% higher. An action essentially includes many postures during its performance; therefore, a posture belonging to an action might be seen in some frames comprising other actions. For instance, the standing posture is sometimes captured at the

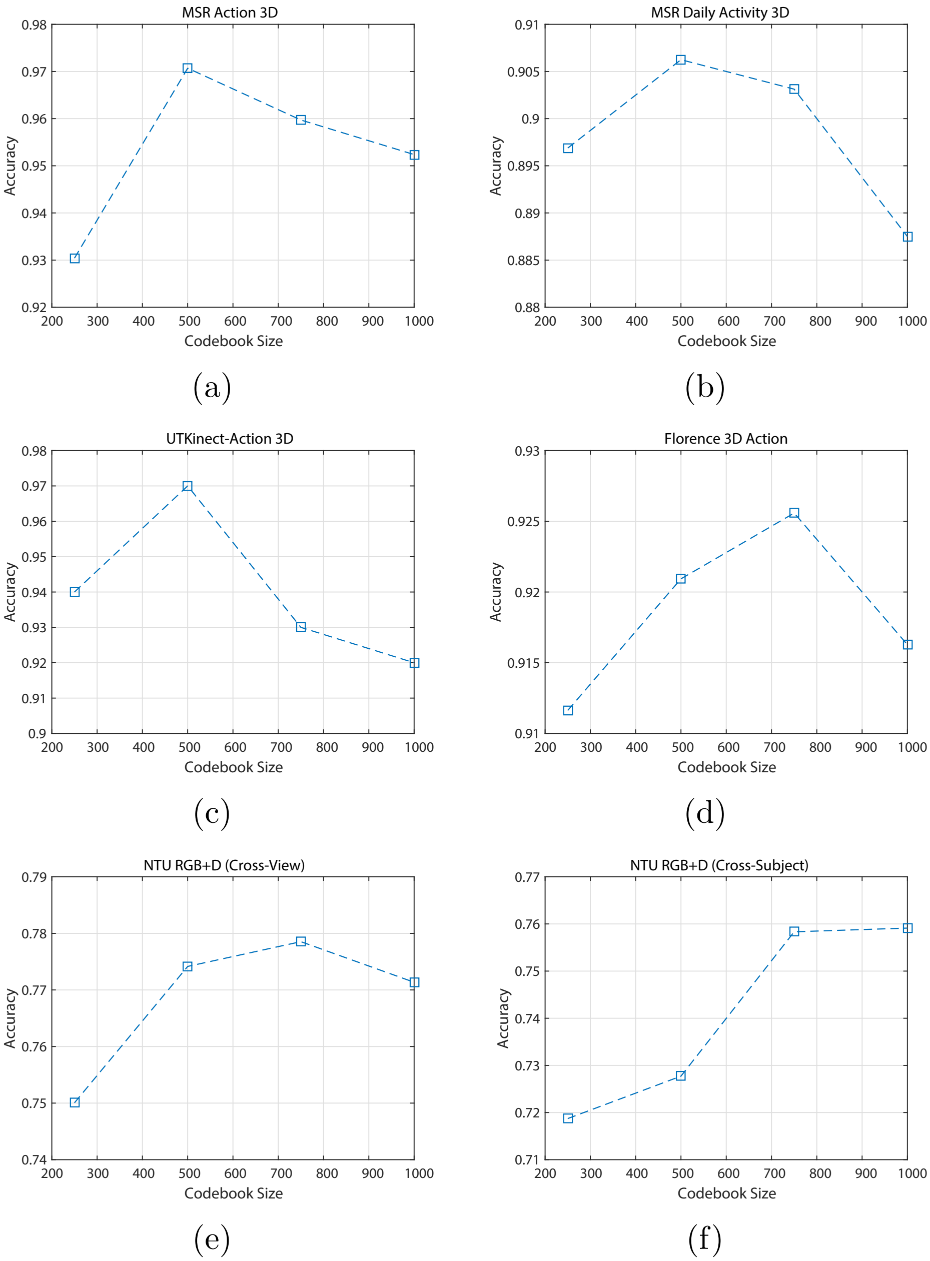
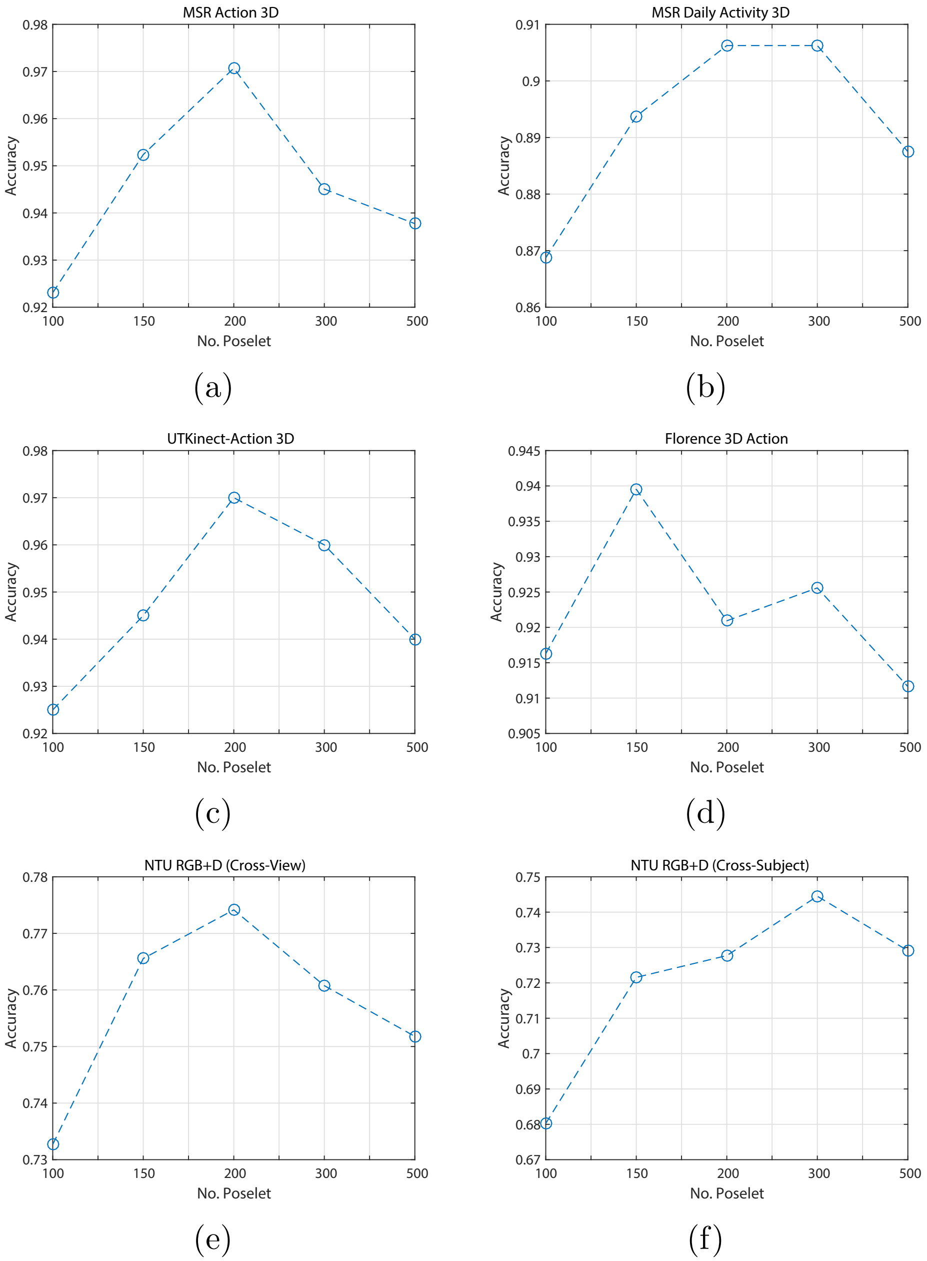


Fig. 3. Evaluation of method sensitivity over the codebook size impact.

beginning of various actions in MSR Action 3D and Florence 3D Action. This creates some confusion in action recognition. However, according to the pose movement explanation between two consecutive frames, the transition feature is more valuable and guaranteed for better action distinction. Moreover, it is recognized that the transition feature is extra efficient with complex datasets, such as MSR Daily Activity 3D and NTU RGB+D Action Recognition.

Codebook size: Besides the type of feature we use, codebook size also has an influence on the overall recognition accuracy, where the recognition results are graphically reported in Fig. 3 with $k = \{250, 500, 750, 1000\}$. It is clear that the accuracy rises on most of the testing datasets following the incremental increase in the codebook size. In particular, the accuracy is significantly improved by 2.41%, 4.03%, and 3.00% on NTU RGB+D Action Recognition (Cross-View), MSR Action 3D, and UTKinect-Action 3D, respectively, when the number of unique codewords is extended from 250 to 500. However, the results are slightly enhanced and even degraded by approximately 4.00% on UTKinect-Action 3D when tuning 750 clusters. As mentioned before, if the size of the codebook is too large, the overall recognition accuracy is not guaranteed to be high due to an overfitting issue. Using $k = 1000$ for such small and simple datasets as MSR Action 3D, UTKinect-Action



(a) (b)

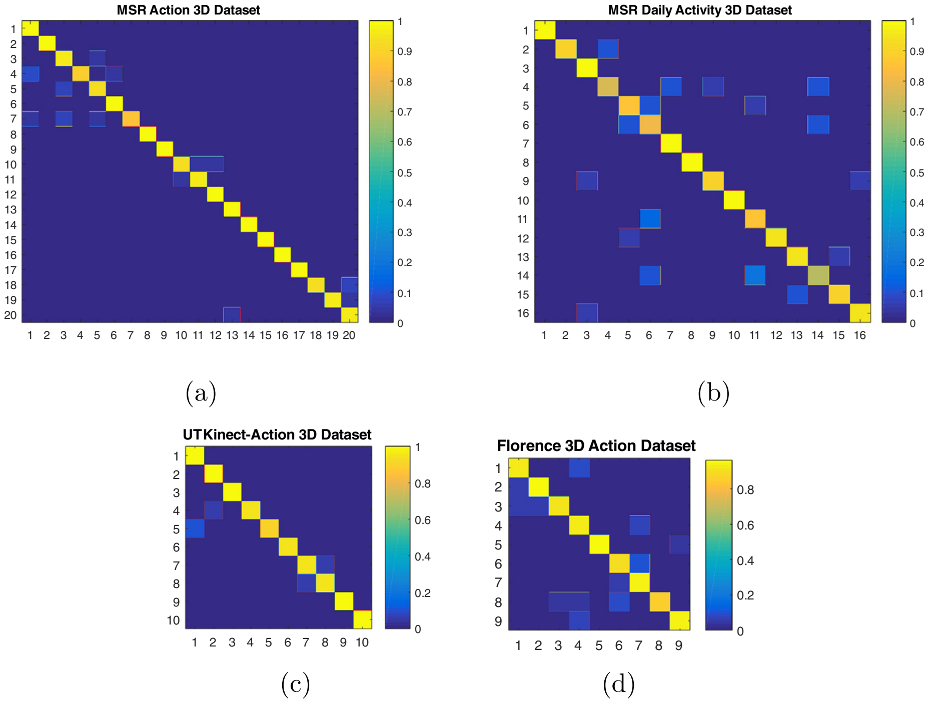
(c) (d)

(e) (f)

Fig. 4. Evaluation of method sensitivity over the number of poselets.

3D, and Florence 3D Action potentially diverges encoding results that lead to recognition confusion. Additionally, the computational cost rapidly grows up following the incremental increase of the number of clusters.

Number of poselets: Playing an important role, the numbers of topics m and subtopics n in PAM have to be initially and suitably chosen. Since the number of topics is characterized to be the number of action classes for each particular dataset, only the number of subtopics corresponding to the number of poselets is determined and evaluated for the overall recognition accuracy. The experimental results on various numbers of poselets are graphically plotted in Fig. 4. When increasing n from 100 to 200, the accuracy is mostly improved for the testing datasets, except Florence 3D Action Recognition; the average improvements reach 2.86% at $k = 150$ and 0.87% at $k = 200$. While the recognition performance continues to increase on MSR Daily Activity 3D and NTU RGB+D Action Recognition (Cross-Subject), the accuracy rate of the remaining datasets slightly decreases when enlarging the number of poselets from 200 to 300. If $n = 500$ is used, all of the evaluations present the same behavior of accuracy reduction. Additionally, the dataset specification has an influence on selecting a proper value of n, e.g., smaller values for relatively simple datasets and larger values for plentiful datasets.



(a) (b)

(c) (d)

Fig. 5. Confusion matrix: (a) MSR Action 3D, (b) MSR Daily Activity 3D, (c) UT Kinect-Action 3D, and (d) Florence 3D Action.

Table 2
Comparison of performance on MSR Action 3D dataset.

Approach	Input data	Acc. (%)
Structure Streaming Skeleton [49]	Skeleton	81.70
Shape and Motion [10]	Depth	82.10
EigenJoints [42]	Skeleton	83.30
Actionlet Ensemble [37]	Skeleton	88.20
HON4D [25]	Depth	88.89
Rate-Invariant Analysis [1]	Skeleton	89.00
Lie Group [35]	Skeleton	89.48
Latent Attribute [32]	Depth	89.80
Dynamic Time Warping + MRF [13]	Skeleton	90.11
Body Surface Context [31]	Depth	90.36
LM^{3}TL [43]	Skeleton	90.53
Markov Random Field [2]	Skeleton	91.01
Riemannian Manifold [4]	Skeleton	92.10
Hierarchical 3D Kernel [12]	Depth	92.73
Multimodal Multipart Learning [28]	Skeleton	93.10
ScTPM + CS-Mltp [22]	Color+Depth	93.83
Joint Angles Similarity + HOG^2 [24]	Color+Depth	94.84
TriViews [3]	Skeleton+Depth	98.20
Deep CNN [38]	Depth	100.00
PAM + Pose-Transition Feature	Skeleton	**97.07**

It is further recognized that using a large number of poselets is not a good idea because of the decrease in of recognition accuracy and the growth of system complexity. To achieve a reasonable tradeoff between the accuracy and complexity, $n = 200$ is tuned as a recommendation for all testing datasets.

4.3.2. Method comparison

The second experiment compares the performance between our proposed approach and state-of-the-art methods on the five datasets following the predefined benchmark protocols corresponding to each particular dataset (i.e., the cross-subject setting for MSR Action 3D and MSR Daily Activity 3D; the leave-one-out cross-validation for UTKinect-Action 3D and Florence 3D Action; the cross-subject and cross-view settings for NTU RGB+D Action Recognition). The confusion matrices are further reported in Fig. 5, except the NTU RGB+D Action Recognition dataset (poor visualization with 60 action classes).

Table 3
Comparison of performance on MSR Daily Activity 3D dataset.

Approach	Input data	Acc. (%)
Rate-Invariant Analysis [1]	Skeleton	70.00
NBNN Bag-of-Poses [26]	Skeleton	70.00
Body Surface Context [31]	Depth	77.80
Mining Mid-level Feature [39]	Skeleton	78.80
Markov Random Field [2]	Skeleton	78.52
Multimodel Multipart Learning [28]	Skeleton	79.38
BIPOD [46]	Skeleton	79.70
Feature of ToSP [30]	Color+Depth	84.40
Deep CNN [38]	Depth	85.00
CoDe4D [47]	Depth	86.00
Actionlet Ensemble [37]	Skeleton	86.00
TriViews [3]	Skeleton+Depth	88.80
Latent Attribute [32]	Depth	88.80
Structure Preserving Projection [44]	Color+Depth	89.80
ScTPM + CS-Mltp [22]	Color+Depth	90.63
PAM + Pose-Transition Feature	Skeleton	**90.63**

Table 4
Comparison of performance on UTKinect-Action 3D dataset

Approach	Input data	Acc. (%)
Deep CNN [38]	Depth	90.91
HOJ3D + LDA [41]	Skeleton	90.92
Riemannian Manifold [4]	Depth	91.50
Hidden CRF [19]	Color+Depth	92.00
Multilayer LSTM [48]	Skeleton	95.96
ST-LSTM (Tree) + Trust Gate [21]	Skeleton	97.00
Lie Group [35]	Skeleton	97.08
TriViews [3]	Skeleton+Depth	98.00
LM^{3}TL [43]	Skeleton	98.80
PAM + Pose-Transition Feature	Skeleton	**97.00**

MSR Action 3D: The comparison outcomes are reported in Table 2. We achieve 97.07% recognition accuracy, which is lower than TriViews [3] and Deep CNN [38]. TriViews exploits the depth sequences and 3D skeleton data according to the front, side, and top view to boost the recognition accuracy. Concretely, five point- and trajectory-based features extracted from depth sequences and 3D skeleton data are jointly used in the TriViews framework. Deep CNN yields a remarkable outcome, but nevertheless, the complexity of three independent CNNs should be efficiently handled in the case of practical development. Compared to others using only 3D skeleton information, PAM outperforms with 3.97-15.37% higher accuracy on the MSR Action 3D dataset.

MSR Daily Activity 3D: The comparison results of MSR Daily Activity 3D dataset are reported in Table 3 in which the accuracy rate of the proposed approach reaches 90.63%, the top place in the accuracy competition with ScTPM + CS-Mltp [22]. Obviously, the combinations of color and depth information in [22,44] are really good to recognize complex activities containing both single actions and human-object interactions; however, they are much heavier than skeleton data for storage and computation. Deep CNN [38] and TriViews [3], two efficient approaches in the MSR Action 3D dataset evaluation, deliver poorly competitive results with 85.00% and 88.80%, respectively. To be better than other skeleton-based approaches, PAM captures a full correlation between features, poselets, and actions to capably recognize the complex activities (*play guitar, play game*, and *write on paper*), and the activities similar in posture, (*drink - call cellphone, use laptop - write on paper*).

UTKinect-Action 3D: Table 4 reports the recognition results in which three approaches including the Lie Group [35], TriViews [3], ST-LSTM (Tree) + Trust Gate [21], and our proposed approach produce the highest accuracy rates which are over 97.00%. The accuracy is higher by 5.00–6.09% than that of the remaining approaches. The information from depth sequences is useful to boost the accuracy of the TriViews approach to 98.00% because some body joints are incorrectly located in human-object interactions. LM^{3}TL [43], a latent max-margin multitask learning model for maximizing the margin partition between action classes, reaches the greatest accuracy at 98.80% with mid-level latent skelets. The Lie Group approach models human actions as curves and maps them to Lie algebra to strengthen the discrimination between different action classes. Although Deep CNN [38] expends a significant amount of computational resources for three neural networks working on depth sequences, its accuracy is disappointingly the worst at 90.91%.

Florence 3D Action: Based on the comparison results reported in Table 5, our proposed approach reaches the runner-up accuracy among the skeleton-based methods. LM^{3}TL [43] yields the highest recognition rate of approximately 93.43%; however, it lacks a scheme for maintaining margin discrimination for multiple “dropped” frames video. We improve the accuracy of the previous work, denoted PAM + Pose Feature [8], from 90.23% to 92.09% thanks to the valuable information

Table 5
Comparison of performance on Florence 3D Action dataset.

Approach	Input data	Acc. (%)
NBNN Bag-of-Poses [26]	Skeleton	82.00
Riemannian Manifold [4]	Skeleton	87.04
PAM + Pose Feature [8]	Skeleton	90.23
Lie Group [35]	Skeleton	90.88
LM^{3}TL [43]	Skeleton	93.42
PAM + Pose-Transition Feature	Skeleton	**92.09**

Table 6
Comparison of performance on NTU RGB+D Action Recognition dataset.

Approach	Input data	Acc. (%)	
		C-S	C-V
Part-aware LSTM [27]	Skeleton	62.93	70.27
ST-LSTM (Tree) + Trust Gate [21]	Skeleton	69.20	77.70
Multilayer LSTM [48]	Skeleton	70.26	82.39
PAM + Pose-Transition Feature	Skeleton	**72.77**	**77.42**

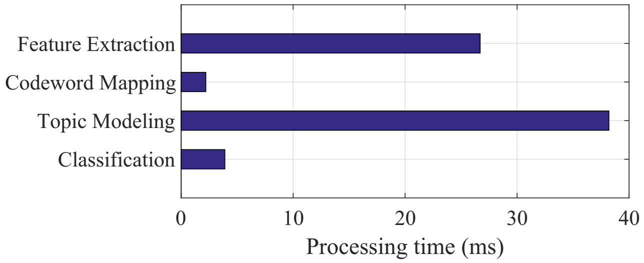


Fig. 6. Average processing time for each testing frame in NTU RGB+D Action Recognition dataset.

coming from the transition feature. Furthermore, the proposed approach performs better than NBNN Bag-of-Poses [26], Riemannian Manifold [4], and Lie Group [35] with 10.09%, 5.05%, and 1.21% higher accuracy, respectively.

NTU RGB+D Action Recognition: As shown in Table 6, the proposed approach defeats Part-aware LSTM [27] in both evaluation protocols, i.e., cross-subject and cross-view with 9.84% and 7.15% greater accuracy, respectively. Compared to Part-aware LSTM [27], ST-LSTM + Trust Gate [21] provides the notable accuracy of 69.20% and 77.70% through the improvements of the gating function and two concurrent domains learning. Based on investigating several geometric features, Zhang et al. [48] achieved the competitive accuracy of 82.39% for the cross-view testing by learning joint-line distances on a three-layer LSTM framework. Due to the challenges of the large numbers of action classes (including single actions, human-object interactions, and human-human interactions) and camera setting configurations (height and distance), the recognition accuracy seems unremarkable with the accuracy of 77.42% in the cross-view protocol. The cross-subject provides less recognition accuracy than the cross-view due to the diversity of the recorded subjects (age, gender, and height).

Based on the experimental results, the proposed method mostly outperforms the state-of-the-art approaches on several testing datasets using only the 3D skeleton data instead of the color and depth information. This specification brings practical benefits, such as storage capacity and computational saving. Another useful specification is the flexibility with varying numbers of joints of each complete skeleton provided by different sources, for instance, 20-joints of Kinect v1 and 25-joints of Kinect v2. Furthermore, the proposed method is capable of processing not only the frame-by-frame but also the frame accumulation schemes by using the window sliding technique.

4.3.3. Computational latency analysis

In most video-based human action recognition systems, low latency is an important impact in addition to high accuracy. This section discusses the computational latency through analyzing the computational complexity. Compared to other datasets, NTU RGB+D Action Recognition is impressive thanks to its diversity of action classes and recording scenario, hence it is reasonable to thoroughly evaluate and analyze the complexity. The performance in terms of processing speed is presented in Fig. 6, where the average timing for each step of our proposed approach is measured by a profiling tool in MATLAB. Considerable time is spent for feature calculation, where the total number of joint distance, joint angle, and joint-plane distance values rapidly increases along with the number of skeleton joints. Therefore, this timing can be reduced by eliminating some minor joints while still retaining recognition accuracy. The timing for mapping codewords from codebook and classification is negligible; however, it can be realized that the mapping time is directly influenced by the dimension of a feature vector. Additionally, it should be noted that the processing time for the NTU RGB+D Action

Recognition dataset is longer than other Kinect v1 based datasets in the feature extraction and codeword mapping steps because more joints have been used. From Fig. 6, topic modeling takes approximately half of the overall processing time for hyper-parameter estimation and probability calculation. With the i7-5700HQ CPU of the notebook computer used for the experiment, the system processes ~ 14 frames per second.

5. Conclusion

In this article, we explore spatio-temporal feature-poselet-action relationships by topic modeling video-based action recognition. We merge joint distance, joint angle, and joint-plane distance extracted in a frame and two consecutive frames to be capable of presenting pose and transition before they are converted to codewords by *k*-means clustering. A set of codewords collected from an action sequence is modeled by our proposed hierarchical model motivated by Pachinko Allocation Model to automatically generate poselet and action assignments. Based on summarizing the correlations among the pose-transition feature as well as the entire associations of feature-poselet-action, our model has the ability to support complicated structures by adopting more realistic assumptions. The proposed approach is evaluated on several well-known 3D datasets under various parameter configurations of the feature type, the codebook size, and the number of poselets to analyze our method sensitivity. Compared to other existing action recognition approaches, we achieve greater recognition accuracy in most of the evaluations, while only using the 3D skeleton information as the input data. We reach remarkable accuracy of 97.07% on MSR Action 3D, 90.63% on MSR Daily Activity 3D, 97.00% on UTKinect-Action 3D, 92.09% on Florence 3D Action, and 77.42% on NTU RGB+D Action Recognition. Our future work will exploit more effective features to robustly handle the cases of human-object and human-human interactions.

References

[1] B.B. Amor, J. Su, A. Srivastava, Action recognition using rate-invariant analysis of skeletal shape trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 1–13.

[2] X. Cai, W. Zhou, L. Wu, J. Luo, H. Li, Effective active skeleton representation for low latency human action recognition, IEEE Trans. Multimedia 18 (2) (2016) 141–154.

[3] W. Chen, G. Guo, Triviews: a general framework to use 3D depth data effectively for action recognition, J. Vis. Commun. Image Represent. 26 (2015) 182–191.

[4] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A.D. Bimbo, 3-D human action recognition by shape analysis of motion trajectories on riemannian manifold, IEEE Trans. Cybern. 45 (7) (2015) 1340–1352.

[5] Y. Du, Y. Fu, L. Wang, Representation learning of temporal dynamics for skeleton-based action recognition, IEEE Trans. Image Process. 25 (7) (2016) 3010–3022.

[6] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with Microsoft Kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.

[7] T. Huynh-The, O. Banos, B.V. Le, D.M. Bui, S. Lee, Y. Yoon, T. Le-Tien, PAM-based flexible generative topic model for 3Dinteractive activity recognition, in: 2015 International Conference on Advanced Technologies for Communications (ATC), 2015, pp. 117–122.

[8] T. Huynh-The, B.-V. Le, S. Lee, Describing body-pose feature - poselet - activity relationship using Pachinko allocation model, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 000040–000045.

[9] T. Huynh-The, B.-V. Le, S. Lee, Y. Yoon, Interactive activity recognition using pose-based spatiotemporal relation features and four-level Pachinko allocation model, Inf. Sci. (Ny) 369 (2016) 317–333.

[10] A. Jalal, S. Kamal, D. Kim, Shape and motion features approach for activity tracking and recognition from Kinect video camera, in: 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, 2015, pp. 445–450.

[11] Y. Kong, Y. Fu, Discriminative relational representation learning for RGB-D action recognition, IEEE Trans. Image Process. 25 (6) (2016) 2856–2865.

[12] Y. Kong, B. Satarboroujeni, Y. Fu, Hierarchical 3D kernel descriptors for action recognition using depth sequences, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1, 2015, pp. 1–6.

[13] S.S. Kruthiventi, R.V. Babu, 3D action recognition by learning sequences of poses, in: Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing, in: ICVGIP '14, 2014, pp. 23:1–23:7.

[14] H. Li, J. Tang, S. Wu, Y. Zhang, S. Lin, Automatic detection and analysis of player action in moving background sports video sequences, IEEE Trans. Circuits Syst. Video Technol. 20 (3) (2010) 351–364.

[15] W. Li, A. McCallum, Pachinko allocation: DAG-structured mixture models of topic correlations, in: Proceedings of the 23rd International Conference on Machine Learning, in: ICML '06, 2006, pp. 577–584.

[16] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 9–14.

[17] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098, doi:10.1109/TPAMI.2015.2400461.

[18] B. Liang, L. Zheng, A survey on human action recognition using depth sensors, in: 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2015, pp. 1–8.

[19] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, Z.-X. Yang, Coupled hidden conditional random fields for RGB-Dhuman action recognition, Signal Process. 112 (2015) 74–82.

[20] A.A. Liu, Y.T. Su, P.P. Jia, Z. Gao, T. Hao, Z.X. Yang, Multiple/single-view human action recognition via part-induced multitask structural learning, IEEE Trans. Cybern. 45 (6) (2015) 1194–1208.

[21] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, in: Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III, Springer International Publishing, 2016.

[22] J. Luo, W. Wang, H. Qi, Spatio-temporal feature extraction and representation for RGB-D human action recognition, Pattern Recognit. Lett. 50 (2014) 139–148.

[23] A. Nava, L. Garrido, R.F. Brena, Recognizing activities using a Kinect skeleton tracking and hidden Markov models, in: 2014 13th Mexican International Conference on Artificial Intelligence, 2014, pp. 82–88.

[24] E. Ohn-Bar, M.M. Trivedi, Joint angles similarities and HOG2 for action recognition, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 465–470.

[25] O. Oreifej, Z. Liu, HON4D: histogram of oriented 4D normals for activity recognition from depth sequences, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 716–723.

[26] L. Seidenari, V. Varano, S. Berretti, A.D. Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 479–485.

[27] A. Shahroudy, J. Liu, T.T. Ng, G. Wang, NTURGB+D: A large scale dataset for 3D human activity analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019.

[28] A. Shahroudy, T.T. Ng, Q. Yang, G. Wang, Multimodal multipart learning for action recognition in depth videos, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2016) 2123–2129.

[29] Y. Shan, Z. Zhang, P. Yang, K. Huang, Adaptive slice representation for human action classification, IEEE Trans. Circuits Syst. Video Technol. 25 (10) (2015) 1624–1636.

[30] Y. Song, S. Liu, J. Tang, Describing trajectory of surface patch for human action recognition on RGB and depth videos, IEEE Signal Process. Lett. 22 (4) (2015) 426–429.

[31] Y. Song, J. Tang, F. Liu, S. Yan, Body surface context: a new robust feature for action recognition from depth videos, IEEE Trans. Circuits Syst. Video Technol. 24 (6) (2014) 952–964.

[32] Y. Su, P. Jia, A. a. Liu, Z. Yang, Discovering latent attributes for human action recognition in depth sequence, Electron. Lett. 50 (20) (2014) 1436–1438.

[33] N.A. Tu, D.-L. Dinh, M.K. Rasel, Y.-K. Lee, Topic modeling and improvement of image representation for large-scale image retrieval, Inf. Sci. (Ny) 366 (2016) 99–120.

[34] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3539–3546.

[35] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.

[36] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1290–1297.

[37] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 914–927.

[38] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P.O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Hum. Mach. Syst. 46 (4) (2016) 498–509.

[39] P. Wang, W. Li, P. Ogunbona, Z. Gao, H. Zhang, Mining mid-level features for action recognition based on effective skeleton representation, in: 2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2014, pp. 1–8.

[40] Z. Wang, L. Li, Q. Huang, Cross-media topic detection with refined CNN based image-dominant topic model, in: Proceedings of the 23rd ACM International Conference on Multimedia, 2015, pp. 1171–1174.

[41] L. Xia, C.C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.

[42] X. Yang, Y. Tian, Effective 3D action recognition using Eigenjoints, J. Vis. Commun. Image Represent. 25 (1) (2014) 2–11.

[43] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, X. Gao, Latent max-margin multitask learning with skelets for 3-D action recognition, IEEE Trans. Cybern. 47 (2) (2017) 439–448.

[44] M. Yu, L. Liu, L. Shao, Structure-preserving binary representations for RGB-Daction recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1651–1664.

[45] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 28–35.

[46] H. Zhang, L.E. Parker, Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction, in: 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 3053–3060.

[47] H. Zhang, L.E. Parker, Code4d: color-depth local spatio-temporal features for human activity recognition from RGB-D videos, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 541–555.

[48] S. Zhang, X. Liu, J. Xiao, On geometric features for skeleton-based action recognition using multilayer LSTM networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 148–157.

[49] X. Zhao, X. Li, C. Pang, Q.Z. Sheng, S. Wang, M. Ye, Structured streaming skeleton – a new feature for online human gesture recognition, ACM Trans. Multimedia Comput. Commun. Appl. 11 (1s) (2014) 22:1–22:18.