Describing Body-Pose Feature - Poselet - Activity Relationship Using Pachinko Allocation Model

Thien Huynh-The *, Ba-Vui Le *, Sungyoung Lee *
* Department of Computer Science and Engineering
Kyung Hee University, Gyeonggi-do, 446-701, Korea
Email: thienht, lebavui, sylee@oslab.khu.ac.kr

***Abstract*—Understanding video-based activities have remained the challenge regardless of efforts from the image processing and artificial intelligence community. However, the rapid developing of computer vision in 3D area has brought an opportunity for the human pose estimation and so far for the activity recognition. In this research, the authors suggest an impressive approach for understanding daily life activities in the indoor using the skeleton information collected from the Microsoft Kinect device. The approach comprises two significant components as the contribution: the pose-based feature extraction under the spatio-temporal relation and the topic model based learning. For extracting feature, the distance between two articulated points and the angle between horizontal axis and joint vector are measured and normalized on each detected body. A codebook is then constructed using the K-means algorithm to encode the merged set of distance and angle. For modeling activities from sparse features, a hierarchical model developed on the Pachinko Allocation Model is proposed to describe the flexible relationship between features - poselets - activities in the temporal dimension. Finally, the activities are classified by using three different state-of-the-art machine learning techniques: Support Vector Machine, K-Nearest Neighbor, and Random Forest. In the experiment, the proposed approach is benchmarked and compared with existing methods in the overall classification accuracy.**

I. Introduction

Nowadays, recognizing human activities has many applications in video surveillance, human - computer interaction, and health care area. Although achieves more impressive results in recent years, human activity recognition is still a quite challenge due to variations of appearance, mutual occlusion, multi-object interaction. Two important considerations for activity recognition are the input sensory data and the activity modeling because they mostly effect to the recognition accuracy and computational cost. Due to typical limitations of RGB videos captured from traditional camera for object detection, pose estimation and activity recognition, the depth camera is therefore considered for this study based on impressive advantages. The complementary information of the depth channel provided by Kinect brings to a realizable solution to solve remaining problems in computer vision [1].

Recent studies proposed activity recognition approaches using the RGB-color videos [2]–[5], however, most of them were restricted by locating body components. In order to evaluate the structural similarity between feature categories, Ryoo et al. [4] designed an impressive spatio-temporal relationship matching algorithm. The algorithm was enable to detect and localize complex non-periodic activities. Meng et al. [2] studied a discriminative model to understand complex activities as the interactions between body components within a person and those between components belongs to different objects. The structural connectivity between objects, human poses, and different body components were validated by a structure search scheme using the max-margin estimation algorithm in Yao's study [5]. By the same way, Kong in [3] proposed a discriminative pattern to model the interactive phrases presenting object motion relationships using the Latent Support Vector Machine (LSVM) technique. Due to neglectfulness of temporal dependencies in phrases and attributes, Kong's approach thereby may confuse different interactions.

Compared with traditional cameras, depth cameras have more advantages in handling illumination changes and provide the extra depth information which motivates and revolutionise for activity detection and recognition research. Besides generating a new dataset of interaction, Multiple Instance Learning (MIL) was proposed by Yun et al. [6] for classification based on the bag of body-component features, such as joint, plane, and velocity features. An efficient descriptor involving an application of a modified Histogram of Oriented Gradient (HOG) algorithm was represented by Ohn-Bar et al. [7] for extracting spatio-temporal features from color and depth images. In Ji's approach [8], interactive activities were modelled by an impressive body model to connect the interactive limbs of different human objects together. Redundant action information was removed by mining essential interactive pairs and poselets for each interaction class. The classification was finally employed using SVM technique with the Radial Basic Function (RBF) kernel on the poselet dictionary. Some fundamental approaches were presented as the combination of 3D-based feature extraction and machine learning techniques, for example as SVM [9], K-Nearest Neighbor (K-NN) [10], Hidden Markov Model (HMM) [11], in which, the body joint coordinate was provided by Kinect sensor. Maximum Entropy Markov Model (MEMM), an advanced technique of HMM, was usually employed for examining recognition accuracy [12]. A hierarchical system as the integration of three different learning techniques including K-NN, SVM, and HMM in the training stage was also recommended by Gaglio et al. [13] to guarantee an acceptable accuracy, real-time processing, and low power consumption. Combining key information from the color channels and the depth channel was studied to amplify the recognition accuracy. In summary, due to natural advantages of Kinect sensor, most of existing 3D-based activity recognition methods has achieved a higher accuracy when compared with other approaches processing on the RGB domain.

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

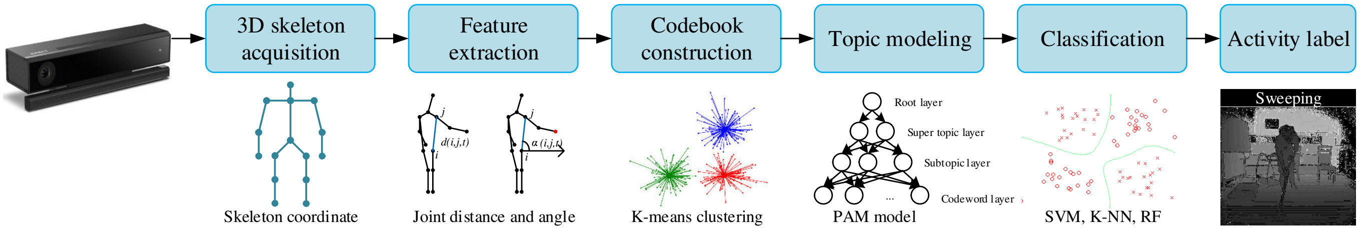


Fig. 1. The workflow of the proposed method. The data collected from Kinect is skeleton coordinate of detected object in the scene.

This paper propose a novel approach to recognize indoor daily life activities using Kinect sensor for collecting skeleton information. First of all, we calculate the distance between two joints and the angle between the joint vector and the horizontal axis on detected body. Extracted features describe the relationships between interactive body components of the detected object in the spatial dimension, the motion of body cannot be therefore captured for deeper understanding activities in the long-term activity. Another challenge in studying activity is similar poselets of different activities. In particular, *reading book* and *calling phone* are two different activities, however, their appearances are mostly similar with the same posture of *sitting*. To overcome those problems, a flexibly hierarchical topic model is developed on the Pachinko Allocation Model (PAM) to exhibit the relationship between the feature - poselet - activity in the multi-frame observation. For preparation of topic modeling, the extracted features are then mapped into codewords using the k-means clustering algorithm. Different with existing approaches mapping each feature, neither distance nor angle as a codeword, we encode merged features as a codeword to emphasize posture discrimination. Due to using the Directed Acyclic Graph (DAG), PAM therefore captures not only correlations among features, but also correlations among poselets and activities. In the final stage, one technique among SVM, K-Nearest Neighbor (K-NN), and Random Forest (RF) is employed for classification task. Besides proposing activity recognition approach, we furthermore introduce a new dataset comprising eight daily life activities recorded by Kinect sensor.

II. The Methodology

The activity recognition approach proposed in this paper has the main following modules: skeleton acquisition, feature extraction, codebook construction, topic modeling, and classification as shown in Fig. 1.

A. Skeleton Acquisition

Although Kinect sensor provides 25 joints on each body as illustration shown in Fig. 2 [1], only 15 key joints have been selected in several common datasets, such as, Florence 3D actions [14] and SBU Kinect Interaction Dataset [6] to reduce the information capacity without accuracy decrement. Some less important joints in action recognition can be omitted: spinebase, neck, handleft, handright, footleft, footright, handtipleft, handtipright, thumbleft, thumbright. Kinect for Windows v2 fundamentally supports to track depth channel and body-frame channel at 30 fps with 512 × 424 of resolution.

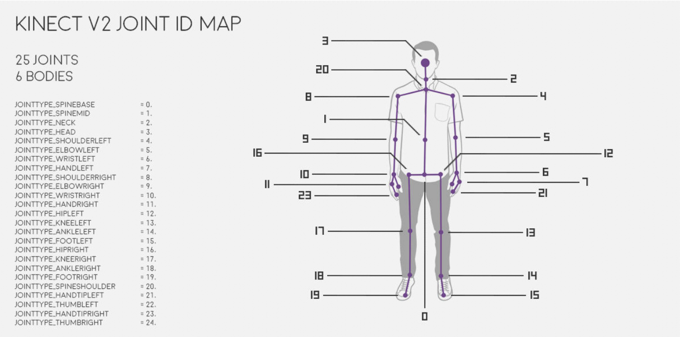

KINECT V2 JOINT ID MAP
25 JOINTS
6 BODIES
JOINTTYPE_SPINEBASE = 0.
JOINTTYPE_SPINEMID = 1.
JOINTTYPE_NECK = 2.
JOINTTYPE_HEAD = 3.
JOINTTYPE_SHOULDERLEFT = 4.
JOINTTYPE_ELBOWLEFT = 5.
JOINTTYPE_WRISTLEFT = 6.
JOINTTYPE_HANDLEFT = 7.
JOINTTYPE_SHOULDERRIGHT = 8.
JOINTTYPE_ELBOWRIGHT = 9.
JOINTTYPE_WRISTRIGHT = 10.
JOINTTYPE_HANDRIGHT = 11.
JOINTTYPE_HIPLEFT = 12.
JOINTTYPE_KNEELEFT = 13.
JOINTTYPE_ANKLELEFT = 14.
JOINTTYPE_FOOTLEFT = 15.
JOINTTYPE_HIPRIGHT = 16.
JOINTTYPE_KNEERIGHT = 17.
JOINTTYPE_ANKLERIGHT = 18.
JOINTTYPE_FOOTRIGHT = 19.
JOINTTYPE_SPINESHOULDER = 20.
JOINTTYPE_HANDTIPLEFT = 21.
JOINTTYPE_THUMBLEFT = 22.
JOINTTYPE_HANDTIPRIGHT = 23.
JOINTTYPE_THUMBRIGHT = 24.
3
20
8
1
9
16
10
11
23
17
18
19
0
2
4
5
12
6
7
21
13
14
15

Fig. 2. A 25-joint skeleton for each body provided by Kinect sensor.

B. Feature Extraction

From the skeleton coordinate dataset, we describe an object appearance by the distance between two joints and the angle metric between the horizontal axis and the joint vector.

Joint distance: The feature is defined as the Euclidean distance between two arbitrary joints and calculated as follows:

$$\begin{aligned} d\left(i,j,t\right) &= \left\| p_{i,t} - p_{j,t} \right\| \\ &= \sqrt{\left(x_{i,t} - x_{j,t}\right)^2 + \left(y_{i,t} - y_{j,t}\right)^2 + \left(z_{i,t} - z_{j,t}\right)^2} \end{aligned} \tag{1}$$

where $p_{i,t} = \{x_{i,t}, y_{i,t}, z_{i,t}\} \in \Re^3$ are the 3D location coordinates of the joint i and j of a detected body at time $t \in T$ corresponding to t^{th} frame. The distance feature describing translation of joints in two consecutive frames is written as follows:

$$d\left(i,j,t-1,t\right) = \left\| p_{i,t-1} - p_{j,t} \right\| \tag{2}$$

Joint angle: The feature is identified by the angle between the horizontal axis $\overrightarrow{Ox}$ and the joint vector $\overrightarrow{ij}$ in the plane $z = 0$:

$$\alpha\left(i,j,t\right) = \angle\left(\overrightarrow{i_t j_t}, \overrightarrow{Ox}\right) = \tan^{-1}\left(\frac{y_{j,t} - y_{i,t}}{x_{j,t} - x_{i,t}}\right) \tag{3}$$

Similarly the angle feature extracted from two joints when considered in two successive frames is computed by the below equation:

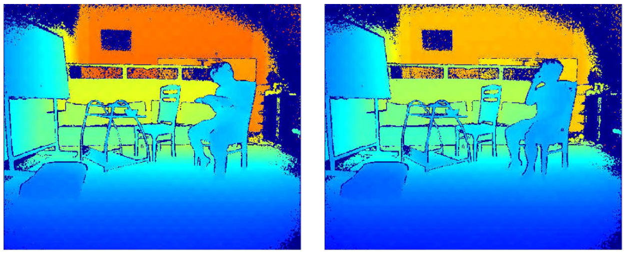
$$\alpha\left(i,j,t-1,t\right) = \angle\left(\overrightarrow{i_{t-1} j_t}, \overrightarrow{Ox}\right) \tag{4}$$

With each pair of two joints, we establish a merged feature consisting of distance and angle:

$$c\left(i,j\right) = \left\{d\left(i,j\right), \alpha\left(i,j\right)\right\} \quad ; i \neq j \tag{5}$$

At the t^{th} frame, totally 210 couple values representing each detected body are calculated from a simplified skeleton model.

[1] https://vvvv.org/documentation/kinect



(a) (b)

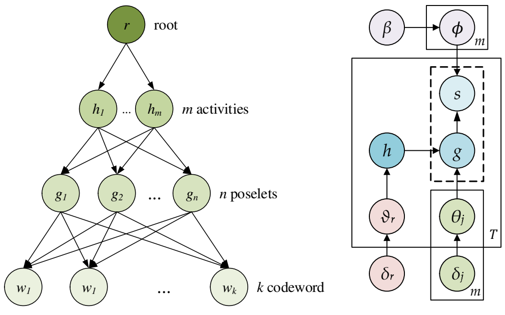
Fig. 3. Samples of two different activities having the similar pose of *sitting*: (a) Reading book, (b) Calling phone.

C. Codebook construction

For the codebook construction, the k-means clustering algorithm using the Euclidean distance metric is applied to cluster the extracted feature dataset. Concretely, an element c in (5) containing information about the joint distance and angle is considered as a codeword, each body is therefore encoded by 210 codewords. In the k-means clustering, the center of each cluster is regarded to be a unique codeword. The parameter k, the number of clusters and also the size of the codebook (the number of vocabulary words) is set in advance. In the training stage, a codebook is built with the k clusters centroid locations. The new features extracted in the testing stage are mapped to codewords by using the constructed codebook. In our proposed method, a particular activity observed in several frames could be offered by the histogram of codewords.

D. Topic modeling

In the previous sections, the spatio-temporal features containing the information of body skeleton within a same frame and two consecutive frames are computed and mapped to codewords. They can be fundamentally used for classification of a short period through common classifiers, however, the long-term studying needs to be discovered for complex activity, such as, *reading book* activity with the pose of *sitting* (see Fig. 3). In this section, the authors therefore suggest a hierarchical model based on the Pachinko Allocation Model to establish the correlation between the sparse feature, individual poselet, and long-term activity. Due to developed on the arbitrary Directed Acyclic Graph, PAM is efficient to learn arbitrary, nested, and possibly sparse activity correlations. PAM is first introduced with arbitrary DAGs by Li and McCallum [15], however the four-level hierarchy structure, a special case with the good tradeoff between complexity and efficiency, consists of one root topic, m super topics at the second level $H = \{h_1, h_2, \ldots, h_m\}$, n subtopics at the third level $G = \{g_1, g_2, \ldots, g_n\}$ and k codewords at the bottom. According to the features comprising the joint distance and angle, codewords generated in the previous stage using the codebook. The super topic and subtopic correspond to the activities and the poselets, respectively. The root is associated with activities, the activities layer are typically associated with poselets, and the poselets are fully connected to the codewords as shown in Fig. 4(a). For each frame, the multinomials of the root and activities are sampled based on the single Dirichlet distribution $\xi_r(\delta_r)$ and $\xi_l(\delta_l)|_{l=1}^{m}$ corresponding to the codewords. The poselets are modelled with multinomial distributions $\phi_{g_l}|_{l=1}^{n}$ which are sampled from the Dirichlet distribution $\xi(\beta)$. Fig. 4(b) represents the graphic model of the four-level PAM version in particular. The notations used in PAM are furthermore summarized in the Table. I. According this model, a frame as a document s in the sequence of T frames $S = \{s_1, s_2, \ldots, s_T\}$, is generated by the following process:



(a) (b)

Fig. 4. Pachinko Allocation Model: (a) Hierarchical topic model (b) Graphic model.

1) Write a multinomial distribution $\vartheta_r^{(s)}$ from a Dirichlet prior $\delta_r^{(s)}$ for frame s.
2) For each activity h_l, write a multinomial distribution $\vartheta_{h_l}^{(s)}$ from $\xi_l(\delta_l)$, where $\vartheta_{h_l}^{(s)}$ is a multinomial distribution over poselets.
3) Write multinomial distributions $\phi_{g_l}|_{l=1}^{n}$ from a Dirichlet prior β for each poselet g_l.
4) For each codeword w in the current frame s:
    - Write an activity $h_{w,s}$ from $\vartheta_r^{(s)}$.
    - Write a poselet $g_{w,s}$ from $\vartheta_{h_{w,s}}^{(s)}$.
    - Write a codeword w from $\phi_{g_{w,s}}$.

Following this process, the joint probability of generating the frame s, the activity assignments $h^{(s)}$, the poselet assignments $g^{(s)}$, and the multinomial distribution $\vartheta^{(s)}$ is calculated as:

$$P\left(s, g^{(s)}, h^{(s)}, \vartheta^{(s)} \middle| \delta, \beta\right) = P\left(\vartheta_r \middle| \delta_r\right) \prod_{l=1}^{m} P\left(\vartheta_{h_l}^{(s)} \middle| \delta_l\right)$$
$$\prod_{w} \left(P\left(h_w \middle| \vartheta_r^{(s)}\right) P\left(g_w \middle| \vartheta_{g_w}^{(s)}\right) P\left(w \middle| \phi_{g_w}\right)\right) \tag{6}$$

Integrating out $\vartheta^{(s)}$ and summing over $h^{(s)}$ and $g^{(s)}$, the marginal probability of a scene is written as follows:

$$P\left(s \middle| \delta, \beta\right) = \int P\left(\vartheta_r^{(s)} \middle| \delta_r\right) \prod_{l=1}^{m} P\left(\vartheta_{h_l}^{(s)} \middle| \delta_l\right)$$
$$\prod_{w} \sum_{h_w, g_w} \left(P\left(h_w \middle| \vartheta_r^{(s)}\right) P\left(g_w \middle| \vartheta_{h_w}^{(s)}\right) P\left(w \middle| \phi_{g_w}\right)\right) d\vartheta^{(s)} \tag{7}$$

The probability of the generated corpus S corresponding to each frame sequence is formulated as follows:

$$P\left(S \middle| \delta, \beta\right) = \int \prod_{l=1}^{n} P\left(\phi_{g_l} \middle| \beta\right) \prod_{s} P\left(s \middle| \delta, \beta\right) d\phi \tag{8}$$

The joint distribution of the corpus S and the topic assignments are given by the below equation:

$$P\left(S, H, G \middle| \delta, \beta\right) = P\left(H \middle| \delta\right) P\left(G \middle| H, \delta\right) P\left(S \middle| G, \beta\right) \tag{9}$$

TABLE I. Notations used in the PAM model

SYMBOL	DESCRIPTION
m	Number of activities
n	Number of poselets
T	Number of frames
k	Number of unique codewords
$\xi_r\left(\delta_r\right)$	Dirichlet distribution associated with the root
$\xi_l\left(\delta_l\right)$	Dirichlet distribution associated with the l^{th} activity
$\xi\left(\beta\right)$	Dirichlet distribution associated with poselet for distance features
$\xi\left(\gamma\right)$	Dirichlet distribution associated with poselet for motion featurres
$\vartheta_r^{(s)}$	The multinomial distribution sampled from $\xi_r\left(\delta_r\right)$ for the root in frame s
$\vartheta_{h_l}^{(s)}$	The multinomial distribution sampled from $\xi_l\left(\delta_l\right)$ for an activity in frame s
ϕ_g	The multinomial distribution sampled from $\xi\left(\beta\right)$ for a poselet g
$h_{w,s}$	The activity h associated with the codeword w in the frame s
$g_{w,s}$	The poselet g associated with the codeword w in the frame s

By integrating out the sampled multinomials, each term is determined as follows:

$$\begin{aligned}
&P\left(H\left|\delta\right.\right)=\int\prod_s P\left(\vartheta_r^{(s)}\left|\delta_r\right.\right)\prod_w P\left(h_w\left|\vartheta_r^{(s)}\right.\right)d\vartheta\\
&P\left(G\left|H,\delta\right.\right)=\int\prod_s\left(\prod_{l=1}^{m}P\left(\vartheta_{h_l}^{(s)}\left|\delta_l\right.\right)\prod_w P\left(g_w\left|\vartheta_{h_w}^{(s)}\right.\right)\right)d\vartheta\\
&P\left(S\left|G,\beta\right.\right)=\int\prod_{l=1}^{n}P\left(\phi_{g_l}\left|\beta\right.\right)\prod_s\left(\prod_w P\left(w\left|\phi_{g_w}\right.\right)\right)d\phi
\end{aligned} \tag{10}$$

The approximate inference result of the condition distribution which samples the super topic and sub-topic assignments for each codeword, is obtained by (11), where $\eta_r^{(s)}$ is the number of occurrences of the root r in the document s; $\eta_l^{(s)}$ is the number of occurrences of activity h_l in the document s; $\eta_u^{(s)}$ is the number of occurrences of poselet g_u in s; $\eta_{lu}^{(s)}$ is the number of times that poselet g_u is sampled from the activity h_l; $\eta_{uv}^{(s)}$ is the number of occurrences of codeword w_v in the poselet h_u. The notation $-w$ indicates activity assignments except word w. The hyper-parameters δ and β can be estimated via the Gibbs sampling algorithm which is detail formulated in [15]. The new data by tagging the joint distance and joint motion features as codewords is generated as the output of PAM. The probability distribution is obtained as the implicit poselet - activity - frame sequence matrix by merging the same codewords in different video contents.

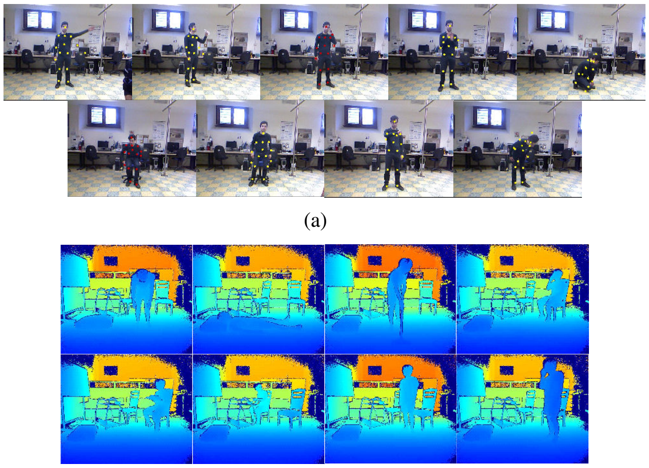
E. Classification

The merged features are viewed as codewords and assigned to a particular poselet and activity patterns by topic modeling. The poselet and activity statistics in every frame sequence are gathered by PAM, then their frequency is observed. Hence, every sequence can be represented by a matrix whose length is the number of poselets and activities. To train the classifier, the labels of vectors and matrices are manually stamped with their classes. In this paper, three standard techniques of Support Vector Machine, K-Nearest Neighbor, and Random Forest are considered for the classification task. Weka 3.6 [16] which is a collection of machine learning algorithms for data mining tasks is used as a tool for performance analysis.

III. Experimental Result

A. Dataset

We validate the proposed method on the Florence 3D Action dataset [14], a well-known dataset widely used for activity recognition benchmark and our dataset.


(a)

(b)

Fig. 5. Two benchmarked datasets: (a) Florence 3D Action dataset shown in color images (*waving*, *drinking from a bottle*, *answering phone*, *clapping*, *tight lacing*, *sitting down*, *standing up*, *reading watch*, and *bowing*), (b) our dataset (*stretching*, *lying down*, *sweeping floor*, *calling phone*, *reading book*, *eating soup*, *watching TV*, and *answering phone*).

Florence 3D Action dataset: Collected at the University of Florence in 2012 and recorded by Kinect device, the dataset has 9 activities: *waving*, *drinking from a bottle*, *answering phone*, *clapping*, *tight lacing*, *sitting down*, *standing up*, *reading watch*, and *bowing* (see Fig. 5a). During acquisition period, 10 objects were asked to act these activities for 2 to 3 times. Totally 215 activity samples, one second for each sample, have been represented in this dataset.

Our dataset: We capture the depth image and skeleton information of 8 common daily life activities using the Kinect for Windows v2: *stretching*, *lying down*, *sweeping floor*, *calling phone*, *reading book*, *eating soup*, *watching TV*, and *answering phone* (see Fig. 5b). From 9 candidates, 72 video sequences are generated with over 90 minutes in the total length.

B. Experimental Setup

In the our proposed method, we extract 210 couple features by (5) consisting of distance and angle for each detected body in the spatio-temporal dimension. For the k-means clustering, couple features are encoded by using 500 codewords as the size of codebook. Fundamentally, greater the number of codewords is, more accuracy the activities are represented by the codebook. However, if the size of codebook is designed too large, the codebook construction process will be very time and memory consuming. In the topic modeling, some parameter have been set up in advance: the number of activity $m = 9$ for Florence 3D Action dataset and $m = 8$ for our dataset, and the number of poselet $n = 100$ for both datasets. Using more poselets in PAM will provide more detail presentation of relationship of codeword and activity, i.e., the accuracy might be improved, however, the greater computational cost will be required for topic modeling. The Dirichlet distribution over activities and posetlets is produced with parameter 0.01. The Gibbs sampling process is performed with 1000 burn-in iterations and then 50 samples are drawn in the following 250 iterations. Feature extraction, codebook construction, and topic modeling are implemented with Matlab 2013a on the desktop

$$P\left(h_w, g_w \middle| S, H_{-w}, G_{-w}, \delta, \beta\right) \propto P\left(w, h_w, g_w \middle| S_{-w}, H_{-w}, G_{-w}, \delta, \beta\right)$$
$$= \frac{P\left(S, H, G \middle| \delta, \beta\right)}{P\left(S, H_{-w}, G_{-w} \middle| \delta, \beta\right)} = \frac{\eta_l^{(s)} + \delta_{rl}}{\eta_r^{(s)} + \sum\limits_{l=1}^{m} \delta_{rl}} \frac{\eta_{lu}^{(s)} + \delta_{lu}}{\eta_l^{(s)} + \sum\limits_{u=1}^{n} \delta_{lu}} \frac{\eta_{uv} + \beta_v}{\eta_u + \sum\limits_{v=1}^{k} \beta_v} \quad (11)$$

computer using CPU Core i5 2.67 GHz and 4 GB RAM. The activity classification is performed by Weka 3.6 with three different machine learning techniques. Similar with Florence 3D Action dataset, we validate our dataset on separating each activity sequence into 1-second samples. All of experiments are fairly benchmarked with 10-fold validation.

C. Result and Discussion

In the first experiment, we investigate the influence of feature on the classification accuracy rate by separating two feature categories: spatial distance-angle and temporal distance-angle. The classification results are presented by confusion matrices in Fig. 6. In two benchmarked datasets, the proposed method reports the greater accuracy using the temporal-based feature instead of the spatial-based feature. Because skeleton information observed at the current frame might be misperceived with other activities also represented by a similar pose, the spatial feature set does not provide the component translation in two consecutive frames. In the Florence 3D Action dataset, *answering phone* has the most confusion with *waving* and *clapping* in classification due to the appearance of *left arm moving* poselet in their activity patterns as the key point. In additions, *sitting down* and *standing* are also confused together because the ending period of *sitting* is the beginning stage of *standing*. In our dataset, *calling phone* and *watching TV* are mostly confused with *reading book* and *answering phone* due to the same poses of *sitting* and *standing*.

In the next experiment, we validate the method with different number of poselet parameter $n = \{50, 100\}$ in PAM. Through the results in Fig. 7, the accuracy improvement when utilizing more poselets in Florence 3D Action dataset is more evident than our dataset, concretely 8.0% versus 0.4%. In order to estimate the effect of classifier on the whole method, we employ three different machine learning techniques. The average results with various parameter configurations and different classifiers are summarized in Table II. Due to the use of Kinect sensor v2, accuracy of joint locating in our dataset is strongly improved from an updated algorithm from Microsoft to explain the greater classification rate when compared with F3D-Action dataset generated by Kinect sensor v1. In this paper, we further compare the proposed method with existing methods consisting of Seidenari et al. [14] and Devanne et al. [17]. Compared with Seidenari's and Devanne's method, the propose scheme outperforms in the average accuracy rate with SVM and RF classifier (see Table III). Seidenari proposed an advanced bag-of-poses approach for recognizing activity from depth camera. Decomposing the body skeleton into a set of kinematic chains permitted to separably align body parts. However, the temporal dynamics of the action information are missed to make the recognition performance less effective. The activity recognition challenge was addressed by Devannce [17] as the problem of calculating the similarity between the trajectory shape in a Riemannian manifold. Some statistical activities, such as lying, sitting, and standing are not appropriate to extract body-part trajectory shapes for matching.

TABLE II. AVERAGE CLASSIFICATION ACCURACY % OF OUR PROPOSED METHOD ON DIFFERENT PARAMETER CONFIGURATIONS

Feature Category	PAM	Classifier	F3D-Action	Our dataset
Spatio Distance-Angle	$n = 100$	K-NN	78.6	94.4
Temporal Distance-Angle	$n = 100$	K-NN	79.3	94.9
Merging Feature	$n = 050$	K-NN	79.0	96.0
Merging Feature	$n = 100$	K-NN	87.0	96.4
Merging Feature	$n = 100$	SVM	89.1	99.2
Merging Feature	$n = 100$	RF	90.2	99.4

TABLE III. COMPARE AVERAGE CLASSIFICATION ACCURACY % OF OUR PROPOSED METHOD WITH EXISTING APPROACHES

Method	F3D-Action	Our dataset
Proposed (K-NN)	87.0	96.4
Proposed (SVM)	89.1	99.2
Proposed (RF)	90.2	99.4
Seidenari et al. [14]	82.3	95.1
Devanne et al. [17]	87.0	97.3

IV. CONCLUSIONS

In this work, we describe a relationship between body-pose feature, poselet, and activity using topic modeling technique for indoor activity recognition. The merged features extracted from the skeleton are encoded to codewords using the k-mean clustering technique before they are modelled by the 4-layer flexible structural model, developed on Pachinko Allocation Model, to automatically produce the poselet and activity model. Due to capturing not only the correlations among features but also the correlations among poselets and activities, the model is successful to provides more expressive power to support complicated structures with adopting more realistic assumptions. We further generate a new dataset including eight daily life activities. Compared with the state-of-art methods in activity recognition area, the proposed method achieve a greater classification accuracy. The method currently contributes to the video-based activity recognition component as one of the most important modules in the Mining Mind [18]. Combining wearable sensor-based [19] and video-based activity recognition will bring an impressive solution for personalized healthcare support. In the future, we will extend more complex and person-person interactive activities with a multi-layer codebook construction technique and advanced PAM.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) NRF-2014R1A2A2A01003914. This work was also supported by the Industrial Core Technology Development Program, funded by the Korean Ministry of Trade, Industry and Energy (MOTIE), under grant number #10049079.

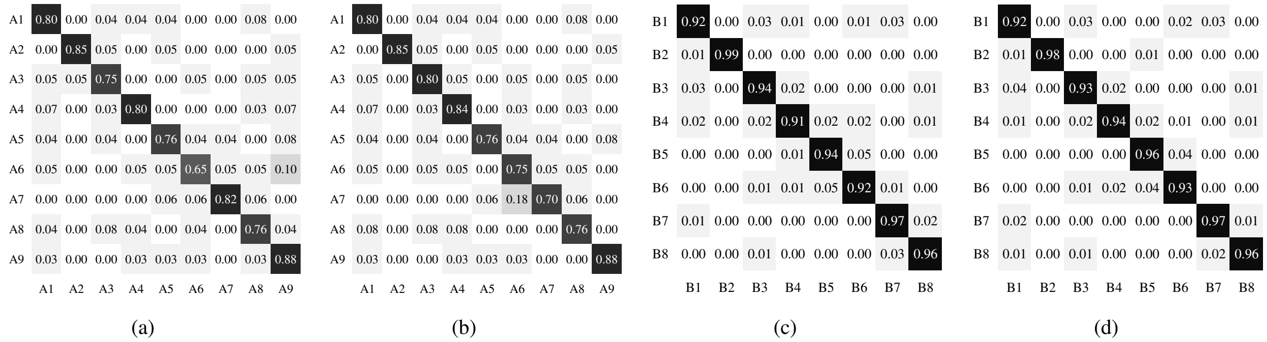


Fig. 6. Confusion matrices of K-NN classifiers on spatial distance-angle and temporal distance-angle category using default parameters in PAM on Florence 3D Action dataset (a, b) and our dataset (c, d). A1-A9 corresponding to 9 activities in Fig. 5a, B1-B9 corresponding to 8 activities in Fig. 5b.

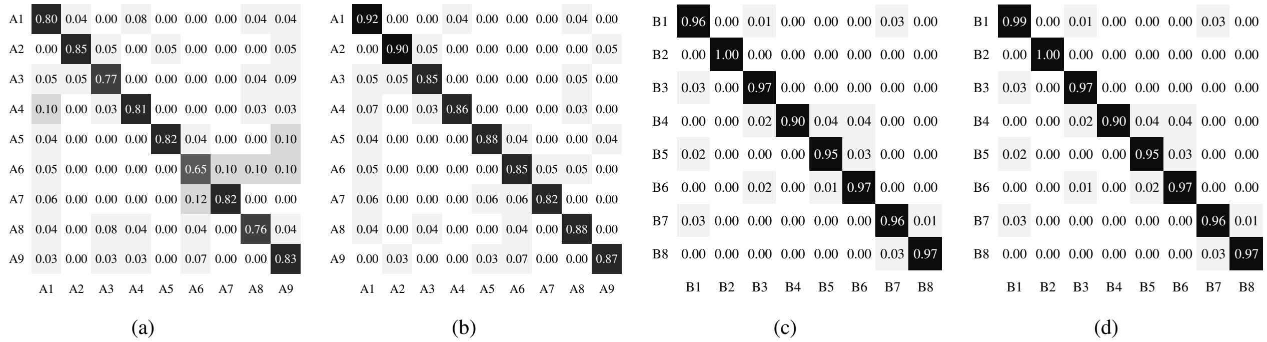


Fig. 7. Confusion matrices of K-NN classifiers on merged feature category using PAM with different numbers of poselet. Results on Florence 3D Action dataset: (a) $n = 50$, (b) $1 = 100$; and results on our dataset: (c) $n = 50$, (d) $n = 100$.

REFERENCES

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1318–1334, Oct 2013.

[2] L. Meng, L. Qing, P. Yang, J. Miao, X. Chen, and D. Metaxas, "Activity recognition based on semantic spatial relation," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Nov 2012, pp. 609–612.

[3] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptionsfor human interaction recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 9, pp. 1775–1788, Sept 2014.

[4] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 1593–1600.

[5] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 17–24.

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

[7] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog2 for action recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 465–470.

[8] Y. Ji, G. Ye, and H. Cheng, "Interactive body part contrast mining for human interaction recognition," in *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, July 2014, pp. 1–6.

[9] Y. Ming, Q. Ruan, and A. Hauptmann, "Activity recognition from rgb-d camera with 3d local spatio-temporal features," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 344–349.

[10] T. Kamizono, H. Abe, K. Baba, S. Takano, and K. Murakami, "Towards activity recognition of learners by kinect," in *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*, Aug 2014, pp. 177–180.

[11] A. Nava, L. Garrido, and R. Brena, "Recognizing activities using a kinect skeleton tracking and hidden markov models," in *Artificial Intelligence (MICAI), 2014 13th Mexican International Conference on*, Nov 2014, pp. 82–88.

[12] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1383–1394, Oct 2013.

[13] S. Gaglio, G. Re, and M. Morana, "Human activity recognition process using 3-d posture data," *Human-Machine Systems, IEEE Transactions on*, vol. 45, no. 5, pp. 586–597, Oct 2015.

[14] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 479–485.

[15] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 577–584.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[17] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *Cybernetics, IEEE Transactions on*, vol. 45, no. 7, pp. 1340–1352, July 2015.

[18] U. C. Laboratory. (2014) Mining mind platform. [Online]. Available: http://www.miningminds.re.kr/english/

[19] L. Chen, J. Hoey, C. Nugent, D. Cook, and Z. Yu, "Sensor-based activity recognition," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 790–808, Nov 2012.