# PAM-Based Flexible Generative Topic Model For 3D Interactive Activity Recognition

Thien Huynh-The *, Oresti Banos *, Ba-Vui Le *, Dinh-Mao Bui *, Sungyoung Lee *, Yongik Yoon †, Thuong Le-Tien ‡
* Department of Computer Engineering
Kyung Hee University, Gyeonggi-do, 446-701, Korea
Email: thienht,oresti,lebavui,sylee@oslab.khu.ac.kr, mao.bui@khu.ac.kr
† Department of Multimedia Science
Sookmyung Women's University, Seoul, 140-172, Korea
Email: yiyoon@sookmyung.ac.kr
‡Department of Electric and Electronics Engineering
Hochiminh University of Technology, Hochiminh City, Vietnam
Email: thuongle@hcmut.edu.vn

***Abstract*—Interactive activity recognition from the RGB videos still remains a challenge, therefore some existing approaches paid the attention to RGB-Depth video process to avoid problems relating to mutual occlusion and redundant human pose and to improve accuracy of skeleton extraction. From the single action to complex interaction activity, it is necessary an efficient model to describe the relationship of body components between multi-human objects. In this research, the authors proposed a hierarchical model based on the Pachinko Allocation Model for interaction recognition. Concretely, the joint features comprising joint distant and joint motion are calculated from the skeleton position and then support to topic modeling. The probabilistic models describing the flexible relationship between features - poselets - activities are generated by this model. Finally, the Binary Tree of Support Vector Machine is applied for classification. Compared with existing state-of-the-arts, the proposed method outperforms in overall classification accuracy (8-21% approximately) with the SBU Kinect Interaction Dataset.**

## I. Introduction

Although receives more attentions from computer vision and artificial intelligence community in recent decades, human activity recognition is still remaining a challenge issue due to variations of appearance, mutual occlusion, multi-object interaction, etc. Previous efforts on human action recognition used the body component motions as input features [1], [2]. Most recent approaches concentrate the collection of low-level features (as local spatial-temporal features) instead of the human body representation (as skeleton) due to limitations of 2-D image processing. In recent years, because of the growth and population of depth sensors, the accuracy of body tracking is adequately improved [3]. However, developing from the single action to the complex activity needs to be considered under the interactive relationship between human objects.

Most recent studies proposed efficient human activity recognition methods on the RGB-color videos [4]–[10], however, their limitations lie on location and separation of body parts, especially when they are overlapped together. In [4], a feature model using a string representation of feature points was proposed to respect the spatio-temporal dynamics of complex activities. In order to measure the structural similarity between sets of feature points, the authors in [8] designed a novel matching algorithm for interaction recognition. An important aspect of a novel mechanism [9] is on the concept of co-occurring feature points of Scale Invariant Feature Transform (SIFT) to describe person-person interactions. One of the most techniques is to examine the interactive body component relationship. In [5], the authors exploited the implicit inter-dependencies existing at both action level and body component level to simplify human interaction recognition. Similarly, complex activities as the interactions between the components of a person (intra-person) and those between the components of different persons (inter-person) were represented by a discriminative model [6]. In [10], the structural connectivity between objects, human pose, and different body components is estimated through a structure search scheme with max-margin estimation algorithm. In the same way, the authors in [7] proposed a discriminative model to encode the interactive phrases describing motion relationships between interacting people based on the latent Support Vector Machine (SVM) formulation. Due to neglectfulness of temporal dependencies in phrases and attributes, the proposed method thereby may confuse different interactions.

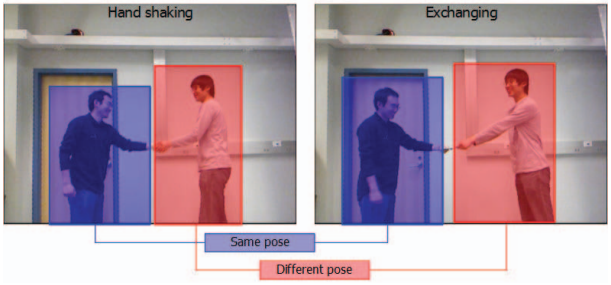



Fig. 1. Interactive activities: hand shaking and exchanging. The blue bounding boxes capture the person with the same pose in different interactions, while the red bounding boxes capture the different poses in different interactions.

Compared with traditional video cameras, RGB-D devices have more advantages in handling illumination changes and

978-1-4673-8374-5/15/$31.00 ©2015 IEEE

provides additional depth information which motivate and revolutionise for the single action detection [11], [12] and complex activity recognition [13]. Determining positions of body joints from a single depth image [14] was represented to forward the single human action recognition. In [13], a new dataset of two-person interactions recorded from inexpensive RGB-D sensor was suggested for evaluation with several geometric relational body-component feature. The Multiple Instance Learning (MIL) algorithm was proposed for classification based on the bag of body-component features, such as joint, plane, and velocity features. A new descriptor involving an application of a modified histogram of oriented gradient (HOG) algorithm was represented in [15] for spatio-temporal feature extraction from color and depth images. An efficient body component model used to connect the interactive limbs of different human objects for interaction representation was proposed in the research [16]. The essential interactive pairs and poselets for each interaction class were also determined to delete redundant action information. The poselets were used to generate the poselet dictionary following bad-of-words to support to SVM classifier with Radial Basis Function (RBF) kernel. The drawback of this method is the poor classification accuracy of complex activities having a high possibility due to similarity of pose interactions, for example with hand shaking and exchanging interactions illustrated as in Fig. 1.

According to contribution in the research [13], the authors in this paper calculate the spatio-temporal poselets for interactive representation using the joint distance and motion features. These poselets describe the relationships between interactive body components of not only the same human object but also the different participants. To overcome the problem of feature co-occurrence in the interaction representation, a flexibly hierarchical model based on the Pachinko Allocation Model (PAM) is proposed to exhibit the relationship between the features - poselets - activity. Concretely, the extracted features are organized into vectors to the codebook construction by using the $k$-means clustering. For topic modeling, a proposed model consists of four levels comprising elements which are fully connected between upper and lower layers. Using the Directed Acyclic Graph (DAG), PAM therefore captures not only correlations among features, but also correlations among poselets and activities. The authors then perform Binary Tree of SVM algorithm for classifying interactive activities.

The rest of the paper is organized as follows: Section 2 describes the proposed method for interactive human activity recognition. The experimental setup, result, and discussion are presented in Section 3. The conclusion is finally stated in Section 4.

## II. The Methodology

The proposed method consists of the following modules: feature extraction, topic modeling, codebook construction, and classification as shown in Fig. 2

### A. Feature extraction

The input data describing the normalized skeleton position of 15 joints per person is provided in the SBU Kinect Interaction Dataset [13]. In order to represent interactive limb pairs, the authors calculate the spatio-temporal joint features

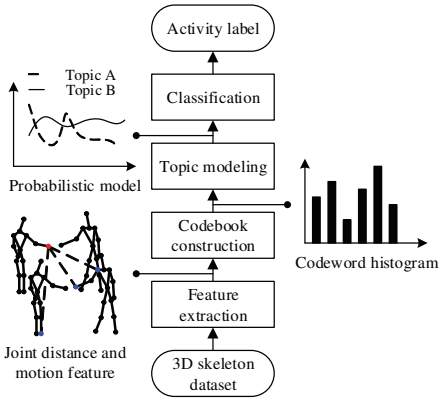



Fig. 2. The workflow of the proposed method.

represented in [13], in which the joints of two persons in a frame, called joint distance, and the joint features of an interactive pair in a set of frames, called joint motion, are identified (see Fig. 3).

**Joint distance**: The joint distance feature is defined as the Euclidean distance between all pairs of joints between two persons in a frame. This parameter captures the distance between two joints in an interaction pose and is calculated as follows:

$$f_d(i,j,t) = \left\| p_{i,t}^x - p_{j,t}^y \right\| \tag{1}$$

where $\left\{p_{i,t}^x, p_{j,t}^y\right\} \in \Re^3$ are the 3D location coordinates of any joint $i$ and $j$ of the human object $x$ and $y$ at the time $t \in T$ corresponding to the $t^{th}$ frame, and this is measured for the same object $(x = y)$ and between two different objects $(x \neq y)$. The joint distance features extracted from an interaction between two objects at the time $t$ are organized into a vector:

$$F_D(t) = \left[ \begin{array}{ccc} F_d^x & F_d^y & F_d^{x,y} \end{array} \right] \tag{2}$$

where $F_d^x = \{f_d(i,j,t) \,|i \in x^t, j \in x^t\}$ and $F_d^y = \{f_d(i,j,t) \,|i \in y^t, j \in y^t\}$ contain the distance between two joints of an object $x$ and $y$, respectively, while $F_d^{x,y} = \{f_d(i,j,t) \,|i \in x^t, j \in y^t\}$ presents a set of joint distance features that are captured between two joints between $x$ and $y$.

**Joint motion**: The joint motion feature is defined as the Euclidean distance between all pairs of joints of two persons in different frames. This feature measures the dynamic motions of interactive limb pairs at time $t - t_0$ and $t$ corresponding to $(t - t_0)^{th}$ and $t^{th}$ frame, and determined as follows:

$$f_m(i,j,t-t_0,t) = \left\| p_{i,t}^x - p_{j,t-t_0}^y \right\| \tag{3}$$

where $t_0$ indicates the time length which is presented by number of frames ($t_0 = 1$ in this research). This feature is measured for the same object $(x = y)$ and for different objects $(x \neq y)$. Similar to the joint distance, the joint motion features are also structured into a vector:

$$F_M(t-t_0,t) = \left[ \begin{array}{cccc} F_m^x & F_m^y & F_m^{x,y} & F_m^{y,x} \end{array} \right] \tag{4}$$

where $F_m^x = \{f_m(i,j,t-t_0,t) \,|i \in x^{t-t_0}, j \in x^t\}$ and $F_m^y = \{f_m(i,j,t-t_0,t) \,|i \in y^{t-t_0}, j \in y^t\}$ contains the distance between two joints of an object $x$ and $y$ at different times, while $F_m^{x,y} = \{f_m(i,j,t-t_0,t) \,|i \in x^{t-t_0}, j \in y^t\}$

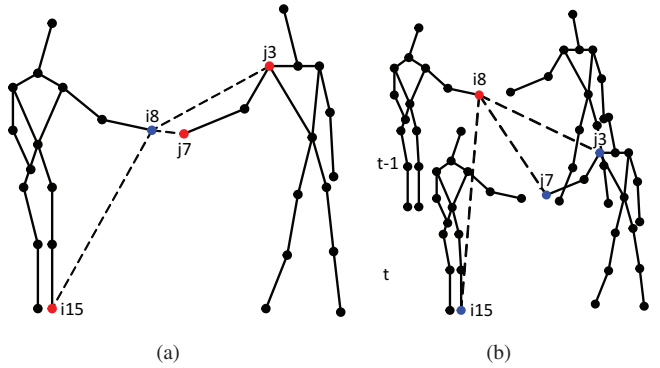



Fig. 3. Body-pose features extracted from skeleton position: (a) Joint distance (b) Joint motion.

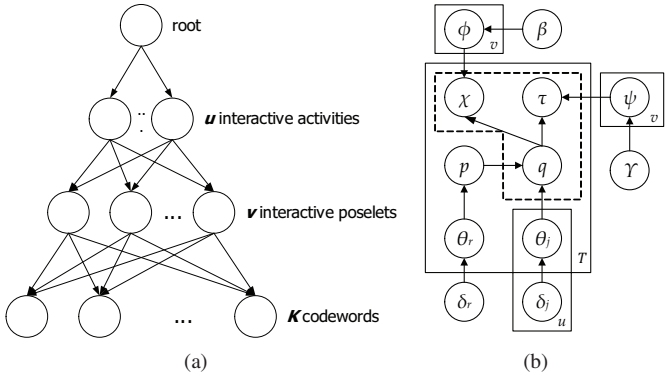



Fig. 4. Pachinko Allocation Model: (a) Hierarchical topic model (b) Graphic model.

and $F_m^{y,x} = \{f_m(i,j,t-t_0,t) \,|i \in y^{t-t_0}, j \in x^t\}$ presents a set of joint motion features that are captured between two joints between $x$ and $y$ at different times.

### B. Codebook construction

For the codebook construction, the authors utilize the $k$-means clustering algorithm based on the Euclidean distance metric to cluster the extracted feature dataset. Concretely, each element $F_d$ in the distance vector $F_D$ or $F_m$ in the motion vector $F_M$ is considered as a codeword. In the $k$-means clustering, the center of each cluster is regarded to be a codeword. The parameter $K$, the number of clusters and also the size of the codebook (the number of vocabulary words) is set in advance. From (2) and (4), there are three and four words which are generated for each joint distance and motion vector. So a set of frames describing an interaction activity can be represented by the histogram of codewords.

### C. Topic modeling

In the previous section, the features describing the joint interaction between objects in the same time and different time are computed and mapped to codewords. Fundamentally, they can been used for interactive action classification of a short period, however, the long time representation needs to be explored. Another issue is the high possibility of different activities comprising more similarly interactive features. This phenomenon will lead to the misclassification, especially with the complex activities, for example as *hand shaking* and *exchanging* as in Fig. 1. Therefore, in this section, the authors proposed a hierarchical model based on the Pachinko Allocation Model to capture the correlation between the interactive features, poselets, and activities. To represent and learn arbitrary, nested, and possibly sparse activity correlations, this model is constructed based on the arbitrary Directed Acyclic Graph.

Although PAM is introduced with arbitrary DAGs, four-level hierarchy structure, a special case [17], consisting of one root topic, $u$ super topics at the second level $\mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_u\}$, $v$ subtopics at the third level $\mathcal{Q} = \{\varrho_1, \varrho_2, \ldots, \varrho_v\}$ and the codewords at the bottom. The codewords are according to the features comprising the joint distance and joint motion which were computed in the previous stage. The super topic and subtopic correspond to the interactive activities and the interactive poselets, respectively. The root is associated to activities, the activities layer are fully associated to interactive poselets, and the poselets are fully connected to the codewords as in Fig. 4(a). The multinomials of the root and activites are sampled for each frame based on a single Dirichlet distribution $g_r(\delta_r)$ and $g_l(\delta_l)|_{l=1}^{u}$ corresponding to the joint distance vector and the joint motion vector, respectively. The poselets are modeled with multinomial distributions $\phi_{\varrho_l}|_{k=1}^{v}$ and $\psi_{\varrho_l}|_{k=1}^{v}$ which are sampled from Dirichlet distribution $g(\beta)$ and $g(\gamma)$. The graphic model for four-level PAM is displayed in Fig. 4(b). The particular notations used in PAM are summarized in the Table. I. According this model, a frame as a document $d$ in the sequence of $T$ frames $\mathcal{D} = \{d_1, d_2, \ldots, d_T\}$, is generated by the following process:

1) Sample a multinomial distribution $\theta_r^{(d)}$ from a Dirichlet prior $\delta_r^{(d)}$ for frame $d$.
2) For each interactive activity $\rho_l$ , sample a multinomial distribution $\theta_{\rho_l}^{(d)}$ from $g_l(\delta_l)$, where $\theta_{\rho_l}^{(d)}$ is a multinomial distribution over interactive poselets.
3) Sample multinomial distributions $\phi_{\varrho_k}|_{k=1}^{v}$ from a Dirichlet prior $\beta$ for each poselet $\varrho_k$.
4) Sample multinomial distributions $\psi_{\varrho_k}|_{k=1}^{v}$ from a Dirichlet prior $\gamma$ for each poselet $\varrho_k$.
5) For each codeword $w$ in the current frame $d$:
   - Sample an interactive activity $\rho_{w,d}$ from $\theta_r^{(d)}$.
   - Sample an interactive poselet $\varrho_{w,d}$ from $\theta_{\rho_{w,d}}^{(d)}$.
   - Sample codeword $w$ from the multinomial $\phi_{\varrho_{w,d}}$ and from the multinomial $\psi_{\varrho_{w,d}}$

Following this process, the joint probability of generating the frame $d$, the interactive activity assignments $\rho^{(d)}$, the interactive posetlet assignments $\varrho^{(d)}$, and the multinomial distribution $\theta^{(d)}$ is calculated as:

$$P\left(d, \varrho^{(d)}, \rho^{(d)}, \theta^{(d)} \middle| \delta, \beta, \gamma\right) = P\left(\theta_r \middle| \delta_r\right) \prod_{l=1}^{u} P\left(\theta_{\rho_l}^{(d)} \middle| \delta_l\right) \prod_{w} \left(P\left(\rho_w \middle| \theta_r^{(d)}\right) P\left(\varrho_w \middle| \theta_{\varrho_w}^{(d)}\right) P\left(w \middle| \phi_{\varrho_w}, \psi_{\varrho_w}\right)\right) \tag{5}$$

Integrating out $\theta^{(d)}$ and summing over $\rho^{(d)}$ and $\varrho^{(d)}$, the

marginal probability of a scene can be calculated as:

$$P(d|\delta,\beta,\gamma) = \int P\left(\theta_r^{(d)}\middle|\delta_r\right)\prod_{l=1}^{u} P\left(\theta_{\rho_l}^{(d)}\middle|\delta_l\right) \prod_{w}\sum_{\rho_w,\varrho_w}\left(P\left(\rho_w|\theta_r^{(d)}\right)P\left(\varrho_w|\theta_{\rho_w}^{(d)}\right)P(w|\phi_{\varrho_w},\psi_{\varrho_w})\right)d\theta^{(d)} \tag{6}$$

The probability of generating the corpus $\mathcal{D}$ corresponding to the overall video is computed by:

$$P(\mathcal{D}|\delta,\beta,\gamma) = \int\prod_{k=1}^{v}\left(P(\phi_{\varrho_k}|\beta)+P(\psi_{\varrho_k}|\gamma)\right) \prod_{d} P(d|\delta,\beta,\gamma)\,d\phi d\psi \tag{7}$$

The joint distribution of the corpus $\mathcal{D}$ and the topic assignments is given by:

$$P(\mathcal{D},\mathcal{P},\mathcal{Q}|\delta,\beta,\gamma) = P(\mathcal{P}|\delta)\,P(\mathcal{Q}|\mathcal{P},\delta)\,P(\mathcal{D}|\mathcal{Q},\beta)\,P(\mathcal{D}|\mathcal{Q},\gamma) \tag{8}$$

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

$$\begin{aligned} P(P|\delta) &= \int\prod_{d} P\left(\theta_r^{(d)}|\delta_r\right)\prod_{w}P\left(\rho_w\middle|\theta_r^{(d)}\right)d\theta \\ P(Q|P,\delta) &= \int\prod_{d}\left(\prod_{l=1}^{u}P\left(\theta_{\rho_l}^{(d)}|\delta_l\right)\prod_{w}P\left(\varrho_w\middle|\theta_{\rho_w}^{(d)}\right)\right)d\theta \\ P(D|Q,\beta) &= \int\prod_{k=1}^{v}P(\phi_{\varrho_k}|\beta)\prod_{d}\left(\prod_{w}P(w|\phi_{\varrho_w})\right)d\phi \\ P(D|Q,\gamma) &= \int\prod_{k=1}^{v}P(\psi_{\varrho_k}|\gamma)\prod_{d}\left(\prod_{w}P(w|\psi_{\varrho_w})\right)d\psi \end{aligned} \tag{9}$$

Finally, the approximate inference result of the condition distribution which samples the super topic and sub-topic assignments for each codeword, is obtained by (6), where $n_r^{(d)}$ is the number of number of occurrences of the root $r$ in the document $d$; $n_l^{(d)}$ is the number of occurrences of activity $\rho_l$ in the document $d$; $n_k^{(d)}$ is the number of occurrences of poselet $\varrho_k$ in $d$; $n_{lk}^{(d)}$ is the number of times that poselet $\varrho_k$ is sampled from the activity $\rho_l$; $n_{kz}^{(d)}$ is the number of occurrences of codeword $w_z$ in the poselet $\varrho_k$. The notation $-w$ indicates activity assignments except word $w$. The hyper-parameters $\delta$, $\beta$, and $\gamma$ can be estimated via the Gibbs sampling algorithm which is described in [17]. The new data by tagging the joint distance and joint motion features as codewords is generated as the output of PAM. By merging the same codewords in different video contents, the probability distribution is obtained as the implicit poselet - activity - frame sequence matrix.

### D. Classification

Joint distance and joint motion features are viewed as codewords and assigned to a particular poselet and activity models by topic modeling. The interactive 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 interactive poselets and activities. To train the classifier, the labels of vectors and matrices are stamped with their classes manually. In this paper, the authors reused the Binary Tree of SVM [18], or BTS for abbreviation, to solve the N-class pattern recognition problem. Each node in the architecture (see Fig. 5) makes binary decision using the original SVM. Based on the recursively dividing the classes into two disjoint groups in every node of the decision tree, the SVM classifier decides the group of unknown sample should be assigned. The class is determined by a clustering algorithm according to the class membership and the interclass distance. In the training phase, BTS has $N-1$ binary classifiers in the best situation ($N$ is the number of classes), while it requires only $\log_{4/3}\left(\frac{N+3}{4}\right)$ binary tests on average when making a decision.

TABLE I. NOTATIONS USED IN THE PAM MODEL

| SYMBOL | DESCRIPTION |
|---|---|
| $u$ | Number of interactive activities |
| $v$ | Number of interactive poselets |
| $T$ | Number of frames |
| $K$ | Number of unique codeworks |
| $g_r\,(\delta_r)$ | Dirichlet distribution associated with the root |
| $g_l\,(\delta_l)$ | Dirichlet distribution associated with the $l^{th}$ activity |
| $g\,(\beta)$ | Dirichlet distribution associated with poselet for distance features |
| $g\,(\gamma)$ | Dirichlet distribution associated with poselet for motion featurres |
| $\theta_r^{(d)}$ | The multinomial distribution sampled from $g_r\,(\delta_r)$ for the root in frame $d$ |
| $\theta_{p_l}^{(d)}$ | The multinomial distribution sampled from $g_l\,(\delta_l)$ for an activity in frame $d$ |
| $\phi_\varrho$ | The multinomial distribution sampled from $g\,(\beta)$ for a poselet $\varrho$ |
| $\psi_\varrho$ | The multinomial distribution sampled from $g\,(\gamma)$ for a poselet $\varrho$ |
| $\rho_{w,d}$ | The interactive activity $\rho$ associated with the codework $w$ in the frame $d$ |
| $\varrho_{w,d}$ | The interactive poselet $\varrho$ associated with the codework $w$ in the frame $d$ |

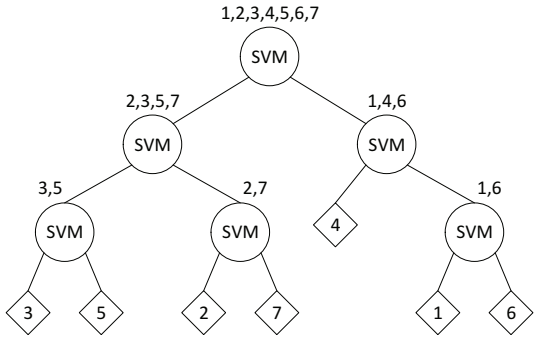


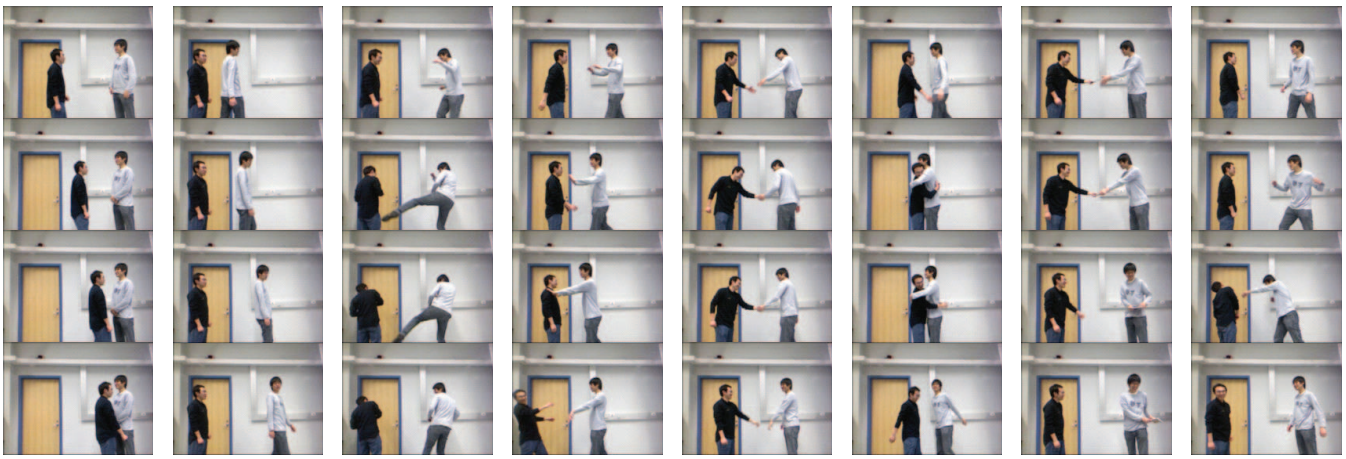

Fig. 5. Illustration of Binary Tree of SVM for a 7-classes sample.

## III. EXPERIMENTAL RESULT

### A. Experimental Setup

The experiments are performed on the SBU Kinect Interaction dataset [13] which comprises of 21 RGB-D video sequence sets, total of 300 interaction videos approximately, describing the interactive activities that are recorded by the Microsoft Kinect Sensor: approaching, departing, pushing, kicking, punching, object exchanging, hugging, and hand shaking as in Fig. 6. The dataset contains RGB images and depth maps with $640 \times 480$ pixels, and also 3D skeleton position of 15 joints. It is important to note that the skeleton in SBU are sometimes not stable on fast and complex motions, especially facing occlusions to lead to fail tracking mission. In the $k$-means clustering, the authors map joint distance and motion features into 500 codewords as the size of codebook. Fundamentally, as the number of codewords increases, the constructed codebooks will represent content of interaction videos more accurately. However,if this parameter is set too large, the process of codebook construction will be very time and memory consuming. In the PAM-based topic modeling, the authors set $u = 8$ interactive activities and $v = 50$ interactive poselets. The Dirichlet distribution over activities and posetlets

$$
\begin{aligned}
P\left(\rho_w, \varrho_w | \mathcal{D}, \mathcal{P}_{-w}, \mathcal{Q}_{-w}, \delta, \beta, \gamma\right) &\propto P\left(w, \rho_w, \varrho_w | \mathcal{D}_{-w}, \mathcal{P}_{-w}, \mathcal{Q}_{-w}, \delta, \beta, \gamma\right) \\
&= \frac{P\left(\mathcal{D}, \mathcal{P}, \mathcal{Q} | \delta, \beta, \gamma\right)}{P\left(\mathcal{D}, \mathcal{P}_{-w}, \mathcal{Q}_{-w} | \delta, \beta, \gamma\right)} \\
&= \frac{n_l^{(d)} + \delta_{rl}}{n_r^{(d)} + \sum_{l=1}^{u} \delta_{rl}} \frac{n_{lk}^{(d)} + \delta_{lk}}{n_l^{(d)} + \sum_{k=1}^{v} \delta_{lk}} \frac{n_{kz} + \beta_z}{n_k + \sum_{z=1}^{K} \beta_z} \frac{n_{kz} + \gamma_z}{n_k + \sum_{z=1}^{K} \gamma_z}
\end{aligned} \tag{6}
$$



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 6. Visualization of the SBU Kinect Interaction dataset: (a) Approaching, (b) Departing, (c) Kicking, (d) Pushing, (e) Hand shaking, (f) Hugging, (g) Exchanging, (h) Punching.


(a) (b) (c) (d)

Fig. 7. Confusion matrices of different body-pose features: (a) Raw data, (b) Joint distance, (c) Joint motion, (d) Joint features. The average classification rates are 69.0%, 82.5%, 89.3%, 90.3%, respectively.

is produced with parameter 0.01. The Gibbs sampling process is performed with 1000 burn-in iterations and then 20 samples are drawn in the following 250 iterations. For BTS classifier, the authors utilizes LibSVM [19] with RBF kernel to solve the binary classification problem. The evaluation is performed by 7-fold cross validation.

### *B. Result and Discussion*

In the experiments, the authors investigate the influence of feature on the classification accuracy rate. Fig. 7 presents the confusion matrices of eight complex interactive activities in the use of different feature types such as the raw data (as skeleton position), joint distance, joint motion, and combined joint feature. The result shows the combined joint feature result in higher classification rate than others, 8% and 21% compared with the joint distance result and raw data approximately, because it describes the relationship between body components of two interactive persons in the spatial and temporal dimension. From the result in Fig. 7(d), *approaching* is mostly confused with *departing* since some interactive poselets of *approaching* have been existed in *departing*, such as *standing*. The *hand shaking* has the most confusion with *exchanging* in classification due to the appearance of *arm stretching* poselet in their interactive activity models. The similar phenomenon occurs with *punching* and *pushing* due to the same reason. Specially, *hugging* has the confusion with the most of other activities, such as *approaching* for beginning period and *departing* for the ending period in the whole of activity.

The authors further compare the proposed method with existing methods, such as Yun et al. [13] and Ji et al. [16],

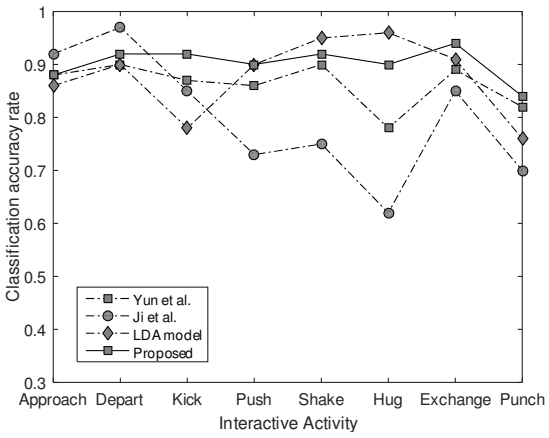



Fig. 8. Classification accuracy comparison of the proposed method with Yun et al. [13], Ji et al. [16], and LDA model [20]. The average classification rates are 90.3%, 80.3%, 86.9%, and 86.3%, respectively.

especially Latent Dirichlet Allocation [20], a well-known generative model. In this experiment, LDA is used for modeling topic likes PAM. The result comparison is reported in Fig. 8. The proposed method presents the higher accuracy rate in average in the competition with published methods. Compared with LDA, a hierarchical model based on PAM technique takes full correlations between features, interactive posetlets to support to the interactive activities.

## IV. Conclusions

In this work, we proposed a hierarchical model for interactive activity recognition. The combined joint feature of joint distance and joint motion are extracted from the skeleton position which is provided from the RGB-D sensor devices. The sparse features are modeled by the 4-layer structural model to automatically generate the interactive poselet and activity model. Due to capturing not only the correlations among features but also the correlations among poselets and activities, the model provides more expressive power to support complicated structures. Compared with the state-of-art methods, the proposed method outperforms in the classification accuracy.

## Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0101-15-1282-00010002, Suspicious pedestrian tracking using multiple fixed cameras). 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.

## References

[1] J. Tian, L. Li, and W. Liu, "A robust framework for 2d human pose tracking with spatial and temporal constraints," in *Digital Image Computing: Techniques and Applications (DICTA), 2014 International Conference on*, Nov 2014, pp. 1–8.

[2] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 489–496.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. IEEE Computer Society, 2011, pp. 1297–1304.

[4] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "A string of feature graphs model for recognition of complex activities in natural videos," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 2595–2602.

[5] Y. Kong and Y. Jia, "A hierarchical model for human interaction recognition," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 1–6.

[6] 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.

[7] 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.

[8] 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.

[9] K. El Houda Slimani, Y. Benezeth, and F. Souami, "Human interaction recognition based on the co-occurrence of visual words," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, June 2014, pp. 461–466.

[10] 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.

[11] C. Zhang and Y. Tian, "Rgb-d camera-based activity analysis," in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, Dec 2012, pp. 1–6.

[12] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, "Discriminative key pose extraction using extended lc-ksvd for action recognition," in *Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on*, Nov 2014, pp. 1–8.

[13] 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, sBU Kinect-Interaction dataset available at http://www3.cs.stonybrook.edu/ kyun/research/kinect_interaction.

[14] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR 11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1297–1304.

[15] 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.

[16] 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.

[17] 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.

[18] B. Fei and J. Liu, "Binary tree of svm: a new fast multiclass training and classification algorithm," *Neural Networks, IEEE Transactions on*, vol. 17, no. 3, pp. 696–704, May 2006.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.