Information Sciences 369 (2016) 317–333

Contents lists available at ScienceDirect

Information Sciences

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

Interactive activity recognition using pose-based spatio–temporal relation features and four-level Pachinko Allocation Model☆

Thien Huynh-The[a], Ba-Vui Le[a], Sungyoung Lee[a,*], Yongik Yoon[b]

[a] Department of Computer Science & Engineering, Kyung Hee University (Global Campus), 1732 Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do, 446–701, Korea
[b] Department of Multimedia Science, Sookmyung Women's University, Cheongpa-ro 47-gil 100 (Cheongpa-dong 2ga), Yongsan-gu, Seoul, 04310, Korea

ARTICLE INFO

Article history:
Received 18 March 2016
Revised 8 June 2016
Accepted 14 June 2016
Available online 25 June 2016

Keywords:
Human interaction
Activity recognition
Four-level PAM

ABSTRACT

In this paper, we go beyond the problem of recognizing video-based human interactive activities. We propose a novel approach that permits to deeply understand complex person-person activities based on the knowledge coming from human pose analysis. The joint coordinates of interactive objects are first located by an efficient human pose estimation algorithm. The relation features consisting of the intra and inter-person features of joint distance and angle, are suggested to use for describing the relationships between body components of the individual persons and the interacting two participants in the spatio-temporal dimension. These features are then provided to the codebook construction process, in which two types of codeword are generated corresponding to distance and angle features. In order to explain the relationships between poses, a flexible hierarchical topic model constructed by four layers is proposed using the Pachinko Allocation Model. The model is able to represent the full correlation between the relation features of body components as codewords, the interactive poselets as subtopics, and the interactive activities as super topics. Discrimination of complex activities presenting similar postures is further obtained by the proposed model. We validate our interaction recognition method on two practical data sets, the BIT-Interaction data set and the UT-Interaction data set. The experimental results demonstrate that the proposed approach outperforms recent interaction recognition approaches in terms of recognition accuracy.

1. Introduction

In recent decades, human activity recognition has been an active research area in computer vision and artificial intelligence due to its wide range of potential applications, such as indoor-outdoor surveillance, human robot interaction, and

☆ 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 research was supported by the MSIP, Korea, under the G-ITRC support program (IITP-2015-R6812-15-0001) supervised by the IITP. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) NRF-2014R1A2A2A01003914.

* Corresponding author. Fax: +82312012514.
E-mail addresses: thienht@oslab.khu.ac.kr (T. Huynh-The), lebavui@oslab.khu.ac.kr (B.-V. Le), sylee@oslab.khu.ac.kr (S. Lee), yiyoon@sookmyung.ac.kr (Y. Yoon).

http://dx.doi.org/10.1016/j.ins.2016.06.016
0020-0255/

human-computer interaction [19,36]. Although it has received attention from the scientist community, an effective method for recognizing activities in the real environment still remains a challenge because of variations of appearance, mutual occlusion, and object interactions. Most existing approaches have concentrated on the low-level features, known as local spatial-temporal features [20,39], instead of the human body representation, known as skeleton, due to limitations in the pose estimation performance. The appearance of limbs strongly varies due to variations in clothing and body shape. Besides, human objects need to be pre-localized and scaled in size by a detector as an initial assumption. However, some notable results in recent years on the human pose estimation have motivated research in human activity recognition [62].

Human activities considered in computer vision can be categorized into two classes: single action and group action. Some approaches were proposed to recognize the activities of one actor as walking, jogging, running, hand waving [56]. The interactive activities between one actor and an object were considered in many recent studies [58]. Some daily life activities in indoor environments can be listed as eating, drinking, typing, and answering phone [33]. Group action, generally performed by visual separable people with complicated interactions, such as walking together, approaching, gathering, has been investigated using human-based features and tracking information for detection and recognition [8,10,45]. Few works handled complex activities of two or more human objects such as hand shaking, hugging, punching, and patting [21,22,60], in which the spatio-temporal relations between two objects are described for activity representation by an interaction model.

In this work, we propose an effective method for human interaction recognition based on a flexible topic model. As a preprocess, locating human articulation is performed by an effective pose estimation algorithm [57]. For representing interactions, the spatio-temporal relation features, calculated from the articulated-pose coordinates, are suggested to use, which include the intra and inter-person features of joint distance and angle. These features describe the relationships between body components of single persons and also interactive participants. To overcome the problem of similar posture interaction representation, we further propose a hierarchical model based on the Pachinko Allocation Model (PAM) to exhibit the relations between features, interactive poselets, and interactions. Concretely, relative features are mapped to visual words by k-means clustering with a constructed codebook. In the topic modeling process, a four-level model which is flexible to connect the upper and lower layers captures the correlations between poselets through codewords and the correlations between interactions through codewords and poselets. Finally, Support Vector Machine (SVM) method is then applied for solving the multi-class classification problem.

The rest of the paper is organized as follows: Section 2 provides discussion on related works. Section 3 describes the proposed method for interaction recognition. The experimental setup, results, and discussion are then presented Section 4. Finally, the conclusions and suggestion for future works are summarized in Section 5.

2. Related work

2.1. Human pose estimation

Human pose estimation, one of the most important stages in the human activity recognition, has received attention in recent years, in which the articulated-pose coordinates or the body part areas in the still images are given. A mostly used technique is the spatial structure coding, often described by the probabilistic graphical model. Although structural-based graphical models allow exact and efficient part inference, they sometimes incorrectly localize body parts in complicated situations. Motivated by the pictorial structure model introduced by Fischler et al. [17], Huttenlocher et al. [16] modeled a human object by a collection of parts arranged in a deformable configuration. By learning latent relationships between different body parts from annotated images, Eichner et al. [13] improved estimation accuracy for unusual poses. From learned appearance models, body parts are ably plugged into any pictorial structure engines. A cascaded model [42] enhances estimation accuracy at different resolutions, where coarse models filter the pose space via max-marginals. Andriluka et al. [2] calculated dense appearance representations using shape context descriptors and then used AdaBoost to train discriminative part classifiers. To obtain the tractability and flexibility, Sapp et al. [41] combined a pictorial structure inference with a non-parametric approach using a subset of model parameters. Furthermore, a shape-based kernel for upper-body pose similarity and a leave-one-out loss function were developed in the learning phase. Building on a successful pictorial structures model, Tian et al. [47] developed an image conditioned model that integrated higher order dependent variables. In recent years, a general and flexible mixture model introduced by Yang et al. [57] based on the standard pictorial structure model captures spatial relations between part positions and co-occurrence relations between part mixtures. Moreover, two novel criteria, the percentage of correct key points (PCK) and the average precision of key points (APK), were proposed to evaluate pose estimation and articulation location, separately and jointly. Two criteria addressed the current shortcomings that are incorrect matching and matching multiple poses to the ground truth. The algorithm has shown notable results of pose estimation with state-of-the-art approaches on practical datasets.

2.2. Features for activity representation

Existing methods for recognizing actions can be categorized into two groups of feature type: local spatio-temporal interest point (STIP) [7,9,12,28,40,43,48,52], and body-pose feature [29,31,49,51,54,56,61]. Activity recognition methods using local features usually develop an effective feature descriptor which provides a pattern or distinct structure found in an

image, such as point, edge, and small image path. Local features extracted in the time domain are usually learned by classification techniques, such as Decision Tree, *K*-Nearest Neighbor, Support Vector Machine, and so on. Human pose based approaches, meanwhile, recognize activities based on the features extracted from articulated joint coordinates, contained by an estimation algorithm, such as distance, angle, velocity, and plane. Local feature based approaches are practically restricted by non-robustness in dynamic scenes, low accuracy with complex actions, and poor relation in interactions. Wu et al. [52] exploited the interest point detector, proposed by Dollar et al.[12], including two separated filters, 2D Gaussian filter and 1D Gabor filter, to produce the high response at each point with significant spatio-temporal intensity variations. Some well-known feature descriptors were usually applied for feature extraction, such as Histogram of Oriented Gradients (HOG) [7,30], Histogram of Optical Flow (HOF) [48], and Scale-Invariant Feature Transform (SIFT) [43], Features From Accelerated Segment Test (FAST) [59], and Motion-Constrained SIFT (MoSIFT) [28]. Based on non-negative matrix factorization (NMF), Eweiwi et al. [14] formulated an informed sampling for action specific regions from interest points to obtain basic flows. In order to avoid outliers from the feature extraction process, Samanta et al. [40] proposed a 3D space-time interest point descriptor using Haar wavelet transform. Motion trajectory providing plentiful spatio-temporal information was potentially utilized for activity recognition [3,37]. A probabilistic trajectory analysis framework [34] was developed for understanding scene activity, in which the trajectory information was clustered into spatial routes and modeled by a hidden Markov model. Compared with STIP-based approaches, pose based methods prove the advantage in complex action and interaction recognition, however, the classification accuracy is sometimes hanged by pose estimation performance. In [31], each action was encoded as a series of synthetic 2D human poses depicted from a wide range of viewpoints. The best matched sequences of actions are then tracked by Viterbi algorithm. Instead of separating pose estimation and action recognition as two individual systems, Yang et al. [56] designed an integrated fashion system that jointly considered poses and actions to directly obtain the pose information. Wang et al. [49] proposed an efficient pose based recognition system, in which the final human articulation was constructed from local parts by a tree structural graphical model. In [51], initial skeletons were collected from a key pose dictionary and particle filters then tracked human upper body parts for activity recognition. A novel feature in [61], namely Poselet Activation Vector, was combined with contextual information, obtained by sparse coding on foreground and background, for action explanation. Extensive pyramidal feature (EPF) constructed from the Gabor filter, Gaussian pyramid, and wavelet transform, was proposed by Liu et al. [29] for pose presentation. The orientation, intensity, and contour information were also encoded by EPFs. A pose dictionary established by shape of contour points from the human silhouette was formulated by Cai et al. [54] to recognize single activities.

2.3. Interactive activity modeling

Another issue, widely considered in the human activity recognition, is interaction modeling from sparse features. This issue is much more importance in the interaction recognition because the inter-relation between objects should be further modeled besides the intra-relation within each object. In [52], a Latent Structure SVM model was introduced for formulating the relationship of action classes – scene classes and the compatibility of multilevel features – action classes. To capture the semantic meaning of body-parts between two interactive objects, Alazrai et al. [1] proposed a motion-pose geometric descriptor (MPGD) based on the concept of anatomical planes. Moreover, a hierarchical framework, consisted of one representation layer and three classification layers, was designed in the recognition phase. Kong et al. [24,25] modeled the actions by large-scale global features and local body part features using a discriminative model to recognize potential interactions. Ryoo et al. [39] introduced a novel spatio-temporal relation matching model to understand human activities captured in their UT-Interaction dataset. A unified-discriminative model was considered for interaction recognition by Meng et al. [32] using inter-person relation features. In [18], modeling activities and matching them in the spatio-temporal dimension were implemented by String of Feature Graph Model. This model is able to recognize activities involving interaction between multiple objects. Recently, topic modeling has been used as an efficient solution for action representation based on visual words, coded from features by vector quantization techniques. Two models, Semilatent Diriclet Allocation (S-LDA) and Semilatent Correlated Topic Model (S-CTM), were suggested for human action recognition in [50]. By pushing the information provided by class labels of training data into these models, the latent topics were matched correctly class labels with quite high accuracy. A novel variant of LDA model [4] including two-layer topics, the mid-level topic describing the local spatial temporal patterns (STPs) and the top-level topic representing mixture distribution of STPs, was proposed by Yang et al. [55] for action group learning. Based on the original LDA [4], Xiao et al. [53] constructed a visual word vocabulary from the STIP and HOG3D features of the cuboids around interest points. A type-2 fuzzy topic model (T2 FTM) [5] was recommended to encode the higher-order uncertainty of each topic from 2D visual words. Although these topic models are impressive in single action recognition, they are sometimes inappropriate and restrictive for interaction because of the compact relations of individual objects and also interactive objects. Mapping directly the features as the visual words to the action as topics might ignore some intermediate states when two poses in an interaction are quite similar together in visualization. In summary, the use of STIP for activity recognition in current approaches could not guarantee the robustness because of STIP's fragility in practical environments. Posture information is really valuable to understand actions in the current frame; however, modeling the articulated-pose features effectively in the time dimension has been not presented in the most of discussed methods. In additions, describing the relations between pose features and actions by some principal modeling techniques cannot provide a deep understanding, especially with the interactive activities. Therefore, we believe that a novel topic model, able to explain the hierarchical relationships of features-postures-activities, is a feasible solution

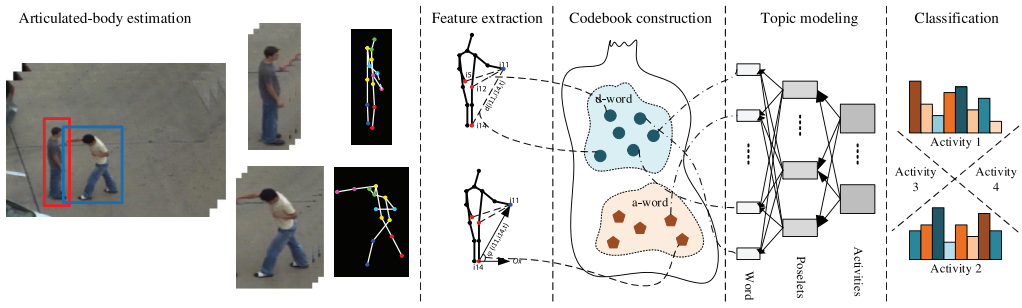

Fig. 1. The workflow of a proposed interaction recognition method using spatio-temporal relation features and topic model. The joint coordinates were achieved by Yang's estimation algorithm on each detected human object. The features used for representing spatio-temporal relation consist of joint distance and angle between pairs of joints. Then the codebook including two types of codewords, *d*-word and *a*-word corresponding to distance and angle feature, was constructed by *k*-mean clustering technique. A hierarchical topic model was suggested for describing the correlation between codeword, poselet, and activity. Finally, interactive activities were classified by a Multi-class SVM.

for remaining limitations. Nevertheless, an expensively computational resource may be needed for the modeling process and becomes a drawback in the comparison with the state-of-the-art methods.

3. Interactive activity recognition method

The proposed interaction recognition method consists of the following modules: articulated-body estimation, spatio-temporal relation feature extraction, codebook construction, topic modeling, and activity classification as Fig. 1.

3.1. Articulated-body estimation

In this work, the authors use an efficient articulated-body estimation algorithm, introduced by Yang et al., to locate the joint coordinates with two patterns of 14-part and 26-part [57] as shown in Fig. 2a–b. Given the bounding box of a human object, key points are evaluated by two criteria, PCK and APK, and then modeled into a tree graphical structure where the nodes of the graph represents body components, and the edges between components illustrate pairwise geometric relations. To detect human objects flexibly and search poses competently in images, a full core function was formulated by associating a compatibility function for part configuration evaluation with a corresponding configuration containing the information of part types and locations. Due to capturing dependence of local appearance on spatial geometry, Yang's model achieved better estimation speed and accuracy if compared with classic articulation models [13,16] using the pictorial structure [17] on the real-life datasets.

In order to boost the performance of articulated joint locating, multiple pose estimators are trained on the testing data sets hereafter to control variance among activities. In the training stage, samples of a particular activity in the Image Parse data set [35] are chosen as positive samples and remainders in the INRIA Person data set [11] are treated as negative samples. The INRIA Person data set comprises non-person images while the Image Parse data set contains 305 pose-annotated images of greatly articulated full-body human poses. In the testing stage, each estimator is tried one by one to select the best result with the maximum score. By this strategy, the estimation accuracy is fairly improved. The data set of 2D coordinates is obtained as the output of this phase.

3.2. Spatio-temporal relation feature extraction

In the interaction case, the human poses should be discriminated on different activities. The pose of individual human object are good enough for single action recognition; however, for interaction recognition, the relations between two objects have to be exploited due to the pose similarity. In the UT-Interaction dataset [39], Punching and Pushing, which present one human object in the standing pose and another in the acting pose should be differentiated through the active poses and the object-object relations. Because most of activities in the practical environments are performed within a time duration, monitoring objects in the time dimension is necessary, in which the temporal features describing the posture translation are extracted. The spatio-temporal relation features are therefore studied from the pose coordinate dataset. The authors calculate the distance of two joints and the angle between a joint vector and the horizontal axis. Fig. 2c–j describes eight feature types extracted from two objects.

Intra-spatio joint distance: The joint distance feature (see Fig. 2c) is defined as the Euclidean distance between a pair of two joints for each human object in a frame; therefore calculated as follows:

$$d_X(i,j,t) = \left\| p^t_{i,X} - p^t_{j,X} \right\| \tag{1}$$

where $p^t_{i,X} \in \Re^2$ is coordinate of joint i belongs to the human object X at the time $t \in T$ corresponding to the tth frame.

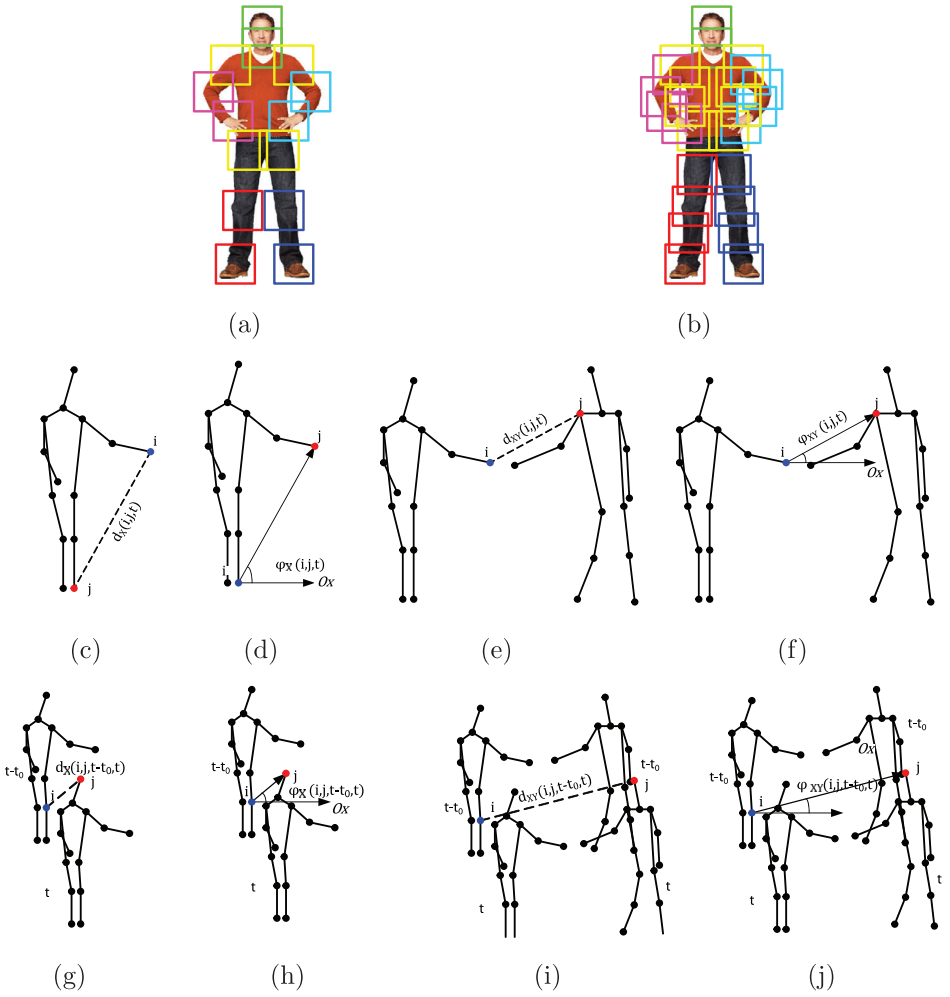

Fig. 2. Illustrations of two articulation patterns: (a) 14-part pattern, (b) 26-part pattern; and the extracted features using the joint coordinate dataset: (c) Intra-spatio joint distance, (d) Intra-spatio joint angle, (e) Inter-spatio joint distance, (f) Inter-spatio joint angle, (g) Intra-temporal joint distance, (h) Intra-temporal joint angle, (i) Inter-temporal joint distance, (j) Inter-temporal joint angle.

Intra-spatio joint angle: The joint angle feature (see Fig. 2d) is defined as the angle between the joint vector $\overrightarrow{p_i p_j}$ and the horizontal axis $\overrightarrow{Ox}$:

$$\varphi_X(i,j,t) = \angle\left(\overrightarrow{p^t_{i,X} p^t_{j,X}}, \overrightarrow{Ox}\right) \tag{2}$$

Inter-spatio joint distance: The inter-spatio joint distance feature (see Fig. 2e) is calculated by Eq. 1, where joints belong to different objects. Particularly, it is measured as follows:

$$d_{XY}(i,j,t) = \left\| p^t_{i,X} - p^t_{j,Y} \right\| \tag{3}$$

where $\left\{p^t_{i,X}, p^t_{j,Y}\right\} \in \Re^2$ are the 2D location coordinates of joint i belongs to the human object X and joint i belongs to the human object Y at the tth frame.

Inter-spatio joint angle: The inter-spatio joint angle feature (see Fig. 2f) is developed from Eq. 2 for two objects:

$$\varphi_{XY}(i,j,t) = \angle\left(\overrightarrow{p^t_{i,X} p^t_{j,Y}}, \overrightarrow{Ox}\right) \tag{4}$$

Intra-temporal joint distance: The intra-temporal joint distance (see Fig. 2g) represents the Euclidean distance between pair of joints belonging to one human object at the current tth frame and the previous $(t-t_0)$th frame:

$$d_X(i,j,t-t_0,t) = \left\| p^{t-t_0}_{i,X} - p^t_{j,X} \right\| \tag{5}$$

where t_0 indicates the time length which is also understood as the number of frames.

Table 1
Category of extracted features.

Feature category	Term	Codebook size
Spatial distance	D_X^S, D_Y^S, D_{XY}^S	K
Temporal distance	$D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$	K
Spatial angle	$\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S$	K
Temporal angle	$\Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$	K
Spatio-temporal distance	$D_X^S, D_Y^S, D_{XY}^S, D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$	K
Spatio-temporal angle	$\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$	K
Spatial distance-angle	$D_X^S, D_Y^S, D_{XY}^S, \Phi_X^S, \Phi_Y^S, \Phi_{XY}^S$	$2K$
Temporal distance-angle	$D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$	$2K$
Merged feature	$D_X^S, D_Y^S, D_{XY}^S, D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$ $\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$	$2K$

Intra-temporal joint angle: The intra-temporal joint angle (see Fig. 2h) describes the angle between the joint vector $\overrightarrow{p_i^{t-t_0} p_j^t}$ and the horizontal axis:

$$\varphi_X(i, j, t-t_0, t) = \angle\left(\overrightarrow{p_{i,X}t - t_0 p_{j,X}^t}, \overrightarrow{Ox}\right) \tag{6}$$

Inter-temporal joint distance: The inter-temporal joint distance (see Fig. 2i) formulates the Euclidean distance between pairs of joints belonging to two different objects at different frames:

$$\begin{aligned} d_{XY}(i, j, t-t_0, t) &= \left\| p_{i,X}^{t-t_0} - p_{j,Y}^t \right\| \\ d_{YX}(i, j, t-t_0, t) &= \left\| p_{i,Y}^{t-t_0} - p_{j,X}^t \right\| \end{aligned} \tag{7}$$

where $d_{XY}(i, j, t-t_0, t)$ is the distance between joint i of the object X at the $(t-t_0)$th frame and joint j of the object Y at the current frame while an opposite case with $d_{YX}(i, j, t-t_0, t)$.

Inter-temporal joint angle: Similarly, the inter-temporal joint angle (see Fig. 2j) expresses the angle features between two different objects in different frames:

$$\begin{aligned} \varphi_{XY}(i, j, t-t_0, t) &= \angle\left(\overrightarrow{p_{i,X}^{t-t_0} p_{i,Y}^t}, \overrightarrow{Ox}\right) \\ \varphi_{YX}(i, j, t-t_0, t) &= \angle\left(\overrightarrow{p_{i,Y}^{t-t_0} p_{i,X}^t}, \overrightarrow{Ox}\right) \end{aligned} \tag{8}$$

Due to the difference in unit, distance and angle features have to be normalized as follows:

$$\widehat{d} = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{9}$$

$$\widehat{\varphi} = \frac{\varphi}{2\pi} \tag{10}$$

All features are obviously summarized and categorized into classes of feature types and dimensions as shown in Table 1, where the terms in categories are identified as follows:

$$\begin{aligned} D_X^S &= \left\{ d(i, j, t) \middle| i \in X^t, j \in X^t, i \neq j \right\} \\ D_Y^S &= \left\{ d(i, j, t) \middle| i \in Y^t, j \in Y^t, i \neq j \right\} \\ D_{XY}^S &= \left\{ d(i, j, t) \middle| i \in X^t, j \in Y^t \right\} \\ D_X^T &= \left\{ d(i, j, t-t_0, t) \middle| i \in Xt - t_0, j \in X^t \right\} \\ D_Y^T &= \left\{ d(i, j, t-t_0, t) \middle| i \in Yt - t_0, j \in Y^t \right\} \\ D_{XY}^T &= \left\{ d(i, j, t-t_0, t) \middle| i \in Xt - t_0, j \in Y^t \right\} \\ D_{YX}^T &= \left\{ d(i, j, t-t_0, t) \middle| i \in Yt - t_0, j \in X^t \right\} \\ \Phi_X^S &= \left\{ \varphi(i, j, t) \middle| i \in X^t, j \in X^t, i \neq j \right\} \\ \Phi_Y^S &= \left\{ \varphi(i, j, t) \middle| i \in Y^t, j \in Y^t, i \neq j \right\} \\ \Phi_{XY}^S &= \left\{ \varphi(i, j, t) \middle| i \in X^t, j \in Y^t \right\} \\ \Phi_X^T &= \left\{ \varphi(i, j, t-t_0, t) \middle| i \in Xt - t_0, j \in X^t \right\} \\ \Phi_Y^T &= \left\{ \varphi(i, j, t-t_0, t) \middle| i \in Yt - t_0, j \in Y^t \right\} \end{aligned}$$

$$\Phi_{XY}^{T} = \left\{ \varphi(i, j, t - t_0, t) \middle| i \in Xt - t_0,\ j \in Y^t \right\}$$
$$\Phi_{YX}^{T} = \left\{ \varphi(i, j, t - t_0, t) \middle| i \in Yt - t_0,\ j \in X^t \right\} \quad (11)$$

Compared with the spatial feature sets only describing the posture relations between two objects in the current frame, the temporal feature sets containing the information of body component translation are predicted to achieve a higher recognition accuracy; however, more computational resources may be needed for calculating temporal features. For example, the temporal distance set has four terms D_X^T, D_Y^T, D_XY^T, and D_YX^T instead of three terms in the spatial distance set as D_X^S, D_Y^S, and D_XY^T. If the 14-part pattern is used, D_X^S contains 91 distance values instead of 196 values in D_X^T. It is necessary to note that the 26-part pattern will rapidly increase the number of features, e.g., 1677 features are extracted from the 26-part pattern instead of 379 values extracted from the 14-part pattern.

3.3. Codebook construction

Some topic models, such as LDA [4] and PAM [27] rely on the existence of a codebook, constructed by a number of visual words. For codebook construction, the authors therefore utilize k-means clustering algorithm based on the Euclidean distance metric to quantize extracted features. The authors separate two types of codeword corresponding to distance and angle. Concretely, each element $\widehat{d}$ in the Spatio–Temporal Distance category is clustered as a distance codeword, denoted d-word, while each element $\widehat{\varphi}$ in the Spatio–Temporal Angle category is considered as an angle codeword, denoted a-word. The parameter K in the clustering algorithm, the number of clusters and also the size of the codebook (the number of vocabulary words), is set in advance for distance and angle feature category. If the method is investigated on the spatial or temporal category class of distance and angle, a mixture codebook of d-word and a-word will be created with $2K$ of size. Meanwhile, a codebook of either d-word or a-word is generated if the distance or angle feature category is utilized (see Table 1). In the testing phase, a frame containing two-object interactions will be represented by a collection of d-word and a-word by mapping from the codebook.

3.4. Four-level Pachinko Allocation Model

In the previous section, the features describing the interaction between two human objects in the spatio-temporal relation are computed and mapped to codewords. Fundamentally, they can be used for interactive action classification of a short period of time, however, the long time representation needs to be explored. Another issue is the high possibility of different activities comprising more similarly interactive features. This unexpected event might lead to the wrongly recognized label, especially with the complex activities, for example as Punching and Pushing. Therefore, in this section, the authors proposed a hierarchical model based on the Pachinko Allocation Model (PAM) to capture the correlation between the relational features, interaction poselets, and interactions. In order to represent and learn arbitrary, nested, and possibly sparse activity correlations, this model is constructed based on the arbitrary Directed Acyclic Graphs (DAGs).

Although PAM is fundamentally introduced with arbitrary DAGs, four-level hierarchy structure, a special case discussed in [27], consists 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 N codewords at the bottom. According to the joint distance and angle features, the codebook comprises d-words and a-words which are described in the previous section. From natural language processing to computer vision, the super topic and subtopic in topic models are corresponding to the interactive activities and the interactive poselets, respectively. The root is associated to interactive activities; the interactive activities are fully connected to interactive poselets; and the interactive poselets are fully linked to the codewords at the bottom of the hierarchical structure as shown in Fig. 3a. The multinomials of the root and activities are sampled for each frame based on a single Dirichlet distribution $g_r(\delta_r)$ and $g_l(\delta_l)|_{l=1}^{u}$, 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 $g(\gamma)$, which will be used for sampling the d-words and a-words in the PAM algorithm. The graphic model for four-level PAM is displayed in Fig. 3b. The particular notations used in PAM are summarized in the Table 2. According this model, a frame s in the sequence of T frames $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$, is generated by the following process:

1. Write a multinomial distribution $\theta_r^{(s)}$ from a Dirichlet prior $\delta_r^{(s)}$ for frame s.
2. For each interactive activity ρ_l, write a multinomial distribution $\theta_{\rho_l}^{(s)}$ over interactive poselets from a Dirichlet distribution $g_l(\delta_l)$, where δ_l is an appropriate Dirichlet prior.
3. Write multinomial distributions $\phi_{\varrho_k}|_{k=1}^{v}$ from a Dirichlet prior β for each interactive poselet ϱ_k.
4. Write multinomial distributions $\psi_{\varrho_k}|_{k=1}^{v}$ from a Dirichlet prior γ for each interactive poselet ϱ_k.
5. For each codeword w in the current frame s:
   - Write an interactive activity $\rho_{w,s}$ from $\theta_r^{(s)}$.
   - Write an interactive poselet $\varrho_{w,s}$ from $\theta_{\rho_{w,s}}^{(s)}$.
   - Write a codeword w from the multinomial $\phi_{\varrho_{w,s}}$ if w is a d-word and from the multinomial $\psi_{\varrho_{w,s}}$ if w is an a-word.

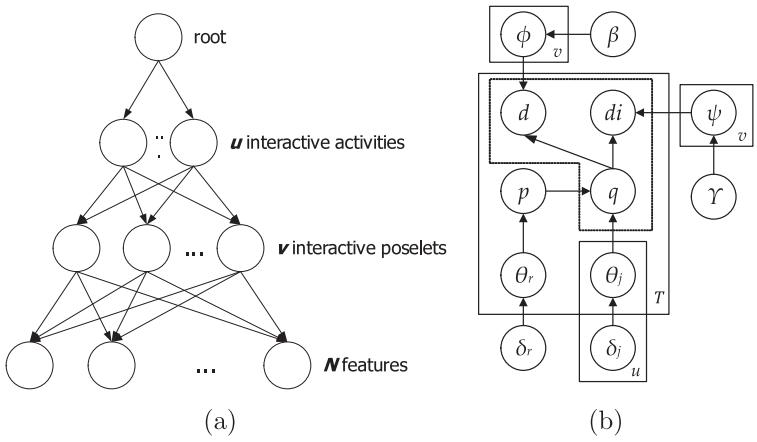

Fig. 3. Pachinko Allocation Model: (a) Hierarchical topic model, (b) Graphic model.

Table 2
Notations used in the PAM model.

Symbol	Description
u	Number of interactive activities
v	Number of interactive poselets
T	Number of frames
N	Number of unique codewords
$g_r(\delta_r)$	Dirichlet distribution associated with the root
$g_l(\delta_l)$	Dirichlet distribution associated with the lth activity
$g(\beta)$	Dirichlet distribution associated with poselet for distance features
$g(\gamma)$	Dirichlet distribution associated with poselet for motion featurres
$\theta_r^{(s)}$	The multinomial distribution sampled from $g_r(\delta_r)$ for the root in frame s
$\theta_{\rho_l}^{(s)}$	The multinomial distribution sampled from $g_l(\delta_l)$ for an activity in frame s
ϕ_{ϱ}	The multinomial distribution sampled from $g(\beta)$ for a poselet ϱ
ψ_{ϱ}	The multinomial distribution sampled from $g(\gamma)$ for a poselet ϱ
$\rho_{w,s}$	The interactive activity ρ associated with the codeword w in the frame s
$\varrho_{w,s}$	The interactive poselet ϱ associated with the codeword w in the frame s

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

$$P\left(s, \varrho(s), \rho(s), \theta^{(s)} \middle| \delta, \beta, \gamma\right) = P(\theta_r|\delta_r) \prod_{l=1}^{u} P\left(\theta_{\rho_l}(s)|\delta_l\right) \times \prod_{w} \left[P\left(\rho_w|\theta_r^{(s)}\right) P\left(\varrho_w|\theta_{\varrho_w}^{(s)}\right) P\left(w|\phi_{\varrho_w}, \psi_{\varrho_w}\right)\right] \tag{12}$$

where $P\left(w|\phi_{\varrho_w}, \psi_{\varrho_w}\right) = P\left(w|\phi_{\varrho_w}\right) P\left(w|\psi_{\varrho_w}\right)$. Integrating out $\theta^{(s)}$ and summing over $\rho^{(s)}$ and $\varrho^{(s)}$, the marginal probability of a frame can be calculated as:

$$P(s|\delta, \beta, \gamma) = \int P\left(\theta_r^{(s)} \middle| \delta_r\right) \prod_{l=1}^{u} P\left(\theta_{\rho_l}^{(s)} \middle| \delta_l\right) \times \prod_{w} \sum_{\rho_w, \varrho_w} \left[P\left(\rho_w|\theta_r^{(s)}\right) P\left(\varrho_w|\theta_{\rho_w}(s)\right) P\left(w|\phi_{\varrho_w}, \psi_{\varrho_w}\right)\right] d\theta^{(s)} \tag{13}$$

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

$$P(\mathcal{S}|\delta, \beta, \gamma) = \int \prod_{k=1}^{v} \left(P\left(\phi_{\varrho_k} \middle| \beta\right) + P\left(\psi_{\varrho_k} \middle| \gamma\right)\right) \prod_{s} P(s|\delta, \beta, \gamma) d\phi d\psi \tag{14}$$

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

$$P(\mathcal{S}, \mathcal{P}, \mathcal{Q}|\delta, \beta, \gamma) = P(\mathcal{P}|\delta) P(\mathcal{Q}|\mathcal{P}, \delta) P(\mathcal{S}|\mathcal{Q}, \beta) P(\mathcal{S}|\mathcal{Q}, \gamma) \tag{15}$$

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

$$
\begin{aligned}
P(\mathcal{P}|\delta) &= \int \prod_{s} P\left(\theta_r^{(s)}|\delta_r\right) \prod_{w} P\left(\rho_w|\theta_r^{(s)}\right) d\theta \\
P(\mathcal{Q}|\mathcal{P},\delta) &= \int \prod_{s} \left(\prod_{l=1}^{u} P\left(\theta_{\rho_l}^{(s)}|\delta_l\right) \prod_{w} P\left(\varrho_w|\theta_{\rho_w}^{(s)}\right) \right) d\theta \\
P(\mathcal{S}|\mathcal{Q},\beta) &= \int \prod_{k=1}^{v} P\left(\phi_{\varrho_k}|\beta\right) \prod_{s} \left(\prod_{w} P\left(w|\phi_{\varrho_w}\right) \right) d\phi \\
P(\mathcal{S}|\mathcal{Q},\gamma) &= \int \prod_{k=1}^{v} P\left(\psi_{\varrho_k}|\gamma\right) \prod_{s} \left(\prod_{w} P\left(w|\psi_{\varrho_w}\right) \right) d\psi
\end{aligned}
\tag{16}
$$

Finally, the approximate inference result of the condition distribution which samples the super topic and subtopic assignments for each codeword, is obtained as follows:

$$
\begin{aligned}
P(\rho_w, \varrho_w|\mathcal{D}, \mathcal{P}_{-w}, \mathcal{Q}_{-w}, \delta, \beta, \gamma) &\propto P(w, \rho_w, \varrho_w|\mathcal{D}_{-w}, \mathcal{P}_{-w}, \mathcal{Q}_{-w}, \delta, \beta, \gamma) \\
&= \frac{P(\mathcal{D}, \mathcal{P}, \mathcal{Q}|\delta, \beta, \gamma)}{P(\mathcal{D}, \mathcal{P}_{-w}, \mathcal{Q}_{-w}|\delta, \beta, \gamma)} \\
&= \frac{n_l^{(s)} + \delta_{rl}}{n_r^{(s)} + \sum_{l=1}^{u} \delta_{rl}} \frac{n_{lk}^{(s)} + \delta_{lk}}{n_l^{(s)} + \sum_{k=1}^{v} \delta_{lk}} \frac{n_{kz} + \beta_z}{n_k + \sum_{z=1}^{K} \beta_z} \frac{n_{kz} + \gamma_z}{n_k + \sum_{z=1}^{K} \gamma_z}
\end{aligned}
\tag{17}
$$

where $n_r^{(s)}$ is the number of occurrences of the root r in the frame s; $n_l^{(s)}$ is the number of occurrences of activity ρ_l in the frame s; $n_k^{(s)}$ is the number of occurrences of poselet ϱ_k in s; $n_{lk}^{(s)}$ is the number of times that poselet ϱ_k is sampled from the activity ρ_l; $n_{kz}^{(s)}$ is the number of occurrences of codeword w_z in poselet ϱ_k. The notation $-w$ indicates the activity assignments except the codeword w. The hyper-parameters δ, β, and γ can be estimated via the Gibbs sampling algorithm which is described in [27]. By tagging the joint distance and joint angle features as codewords, the new data is generated as the output of PAM. The probability distribution is obtained as the implicit poselet – activity – frame sequence matrix from merging the same codewords in different video contents

3.5. Classification

The joint distance and joint angle features are viewed as codewords and assigned to particular interactive poselet and activity models by a four-level Pachinko Allocation Model. The interactive poselet and activity statistics in every frame sequence are gathered by PAM, then their frequency is observed. Hence, every sequence is represented by a matrix whose length is the number of interactive poselets and activities. The interaction recognition is performed on these matrices corresponding to interaction videos. To solve the N-class pattern recognition problem, the authors utilize the Binary Tree of SVM [15], or BTS for abbreviation, in which each node in the tree produces a 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 group of unknown sample will be identified by the SVM classifier. In the training phase, BTS has $N-1$ binary classifiers (N is the number of classes) while it requires only $\log_{4/3}\left(\frac{N+3}{4}\right)$ binary tests on average to make a decision.

4. Experiments and discussions

4.1. Dataset and experiment setup

In this paper, three experiments are performed on two well-known interaction datasets, BIT-Interaction dataset [24] and UT-Interaction dataset [39]. All of the experiments were performed on a desktop PC running Windows 7 Operating System with a 2.67-GHz Intel Core i5 CPU and 4GB of RAM. Matlab 2013a was used to make the simulations. The proposed method was evaluated using the 10-fold cross-validation for all experiments.

BIT-Interaction data set has eight classes of human interactive activity (bowing, boxing, hand shaking, high-five, hugging, kicking, patting, and pushing) as shown in Fig. 4a–h, with 50 short videos ($\sim$ 2 s) per class. Each video is recorded with a resolution of 320 $\times$ 240 and a rate of 30 fps (frame per second). Videos are captured in realistic scenes, included indoor and outdoor environments, with partial occluded body components, dynamic object movements, and different viewpoints in various illumination conditions.

UT-Interaction data set consists of six interactions (hand shaking, hugging, kicking, pointing, punching and pushing) as shown in Fig. 4i–n. Each interaction is presented by 10 videos whose lengths are around 1 min. Totally, there are 60 videos for six classes provided in the data set. Those videos are captured with a resolution of 720 $\times$ 480 and a frame rate of 30 fps with slightly different zoom rate and camera jitter.

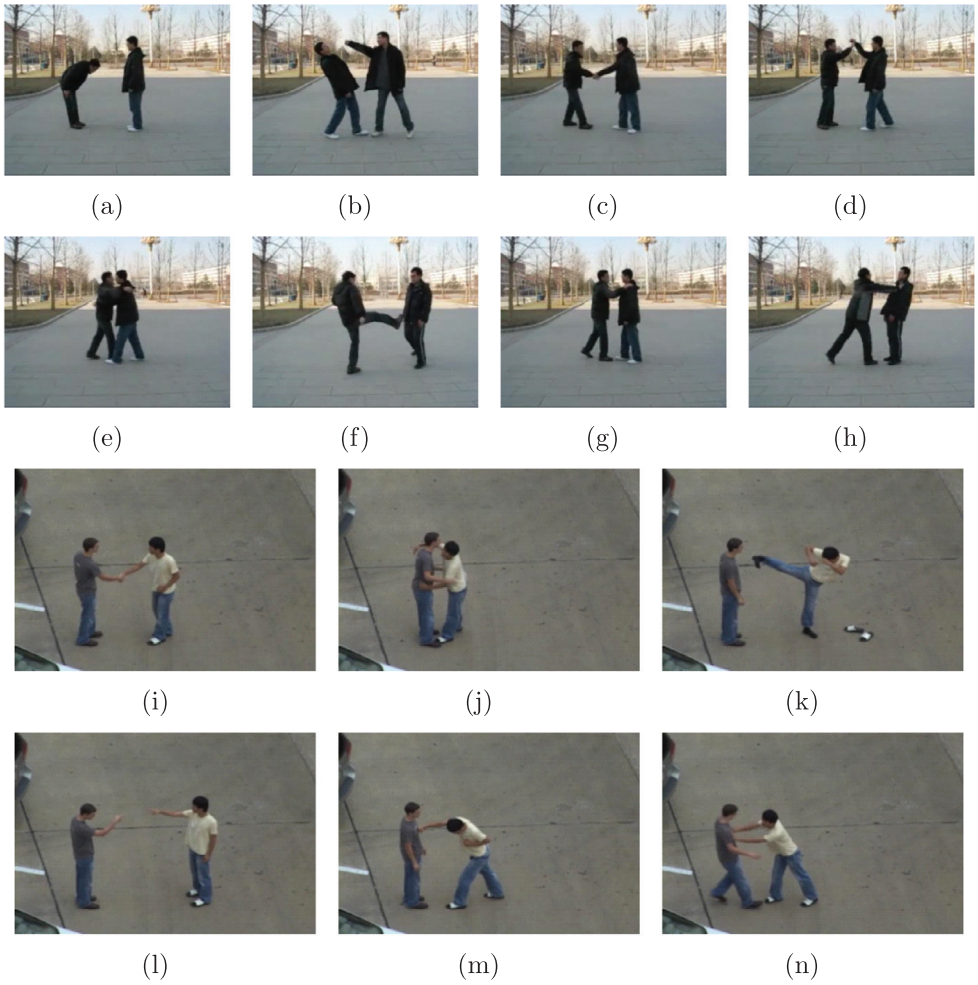

Fig. 4. Eight interactive activities in the BIT-Interaction data set: (a) Bowing, (b) Boxing, (c) Hand shaking, (d) High-five, (e) Hugging, (f) Kicking, (g) Patting, (h) Pushing. Six interactive activities in the UT-Interaction data set: (i) Hand shaking, (j) Hugging, (k) Kicking, (l) Pointing, (m) Punching, (n) Pushing.

In two data sets, the individual human objects in each frame were extracted by their bounding boxes supported in dataset owners for pose estimation. This strategy aim is to improve the estimation accuracy and computational speed because of searching body parts in a segmented area instead of a whole image. Moreover, the authors consider both the 14-part and 26-part patterns using the INRIA Person dataset [11] and the Image Parse data set [35] for pose learning. The parameter configuration for pose estimation algorithm [57] is set up with $\alpha = 0.2$, h and w as the height and width of the bounding box. A codebook with 1000 of size, consisted of 500 d-words and 500 a-words, is constructed using the k-mean clustering algorithm. In the four-level PAM model, the numbers of interactive activities u are defined to 8 and 6 for the BIT-Interaction data set and the UT-Interaction data set, respectively, and the number of interactive poselets v was set to 150 for both of them. The Dirichlet distribution over activities and poselets 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 utilize LibSVM [6] with RBF kernel to solve the multi-class classification problem.

Three experiments are explained in detail as follows:

- In the first experiment, the recognition method is validated for each data set on two articulation estimation patterns. This experiment investigates the influence of pose estimation performance on activity recognition accuracy.
- In the second experiment, the authors validate the feature types, distance and angle feature sets in the spatial and temporal dimension, using the 26-part pattern for pose estimation. This benchmark proves that recognition accuracy also depends on relation features.
- Finally, the proposed PAM-based hierarchical topic model was compared with LDA, a standard topic model, and the state-of-the-art interaction recognition methods on the same testing datasets.

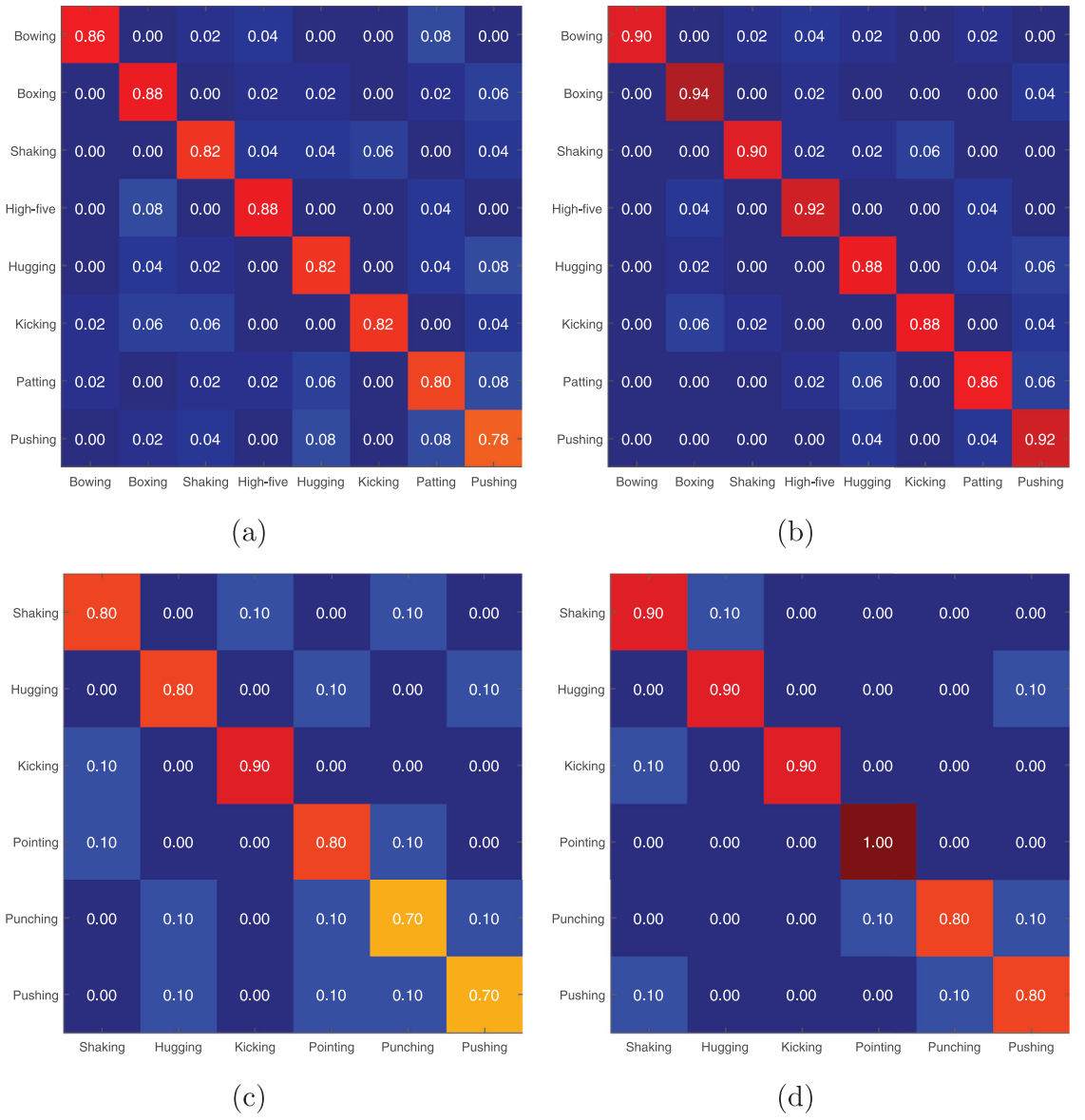

Fig. 5. Confusion matrices of the SVM classifier using spatio-temporal distance feature set on the BIT-Interaction data set: (a) 14-part pattern, (b) 26-part pattern; and on UT-Interaction data set: (c) 14-part pattern, (d) 26-part pattern.

4.2. Experiment results and discussions

The confusion matrices for the first experiment on two interaction datasets are shown in Fig. 5. The feature category used for the estimation pattern test is the spatio-temporal distance set (see Table 1). For the BIT-Interaction dataset (Fig. 5a–b), the 26-part pattern recognizes more accurate than 14-part pattern in most of activity classes. More valuable features which are extracted from the 26-part pattern are useful for the complex activity understanding. An improvement of classification accuracy is also obtained in the case of the UT-Interaction data set (Fig. 5c–d) where the 26-part pattern outperformed the 14-part pattern on five activities among six in total. However, more estimated parts (midway points between limbs, such as mid-upper arm, mid-lower arm, etc) significantly increase the computational cost of estimation and feature extraction processes, for instance with the spatial distance category, $(325 + 2 \times 676)$ features calculated from the 26-part pattern instead of $(91 + 2 \times 144)$ features calculated from the 14-part pattern.

In the second experiment, the authors investigate the proposed method on different feature categories using the 26-part pattern. The classification results on the BIT-Interaction data set are reported by the confusion matrices in Fig. 6 and summarized in Table 3. Totally, there are eight examined categories: the spatial distance set, temporal distance set, spatial

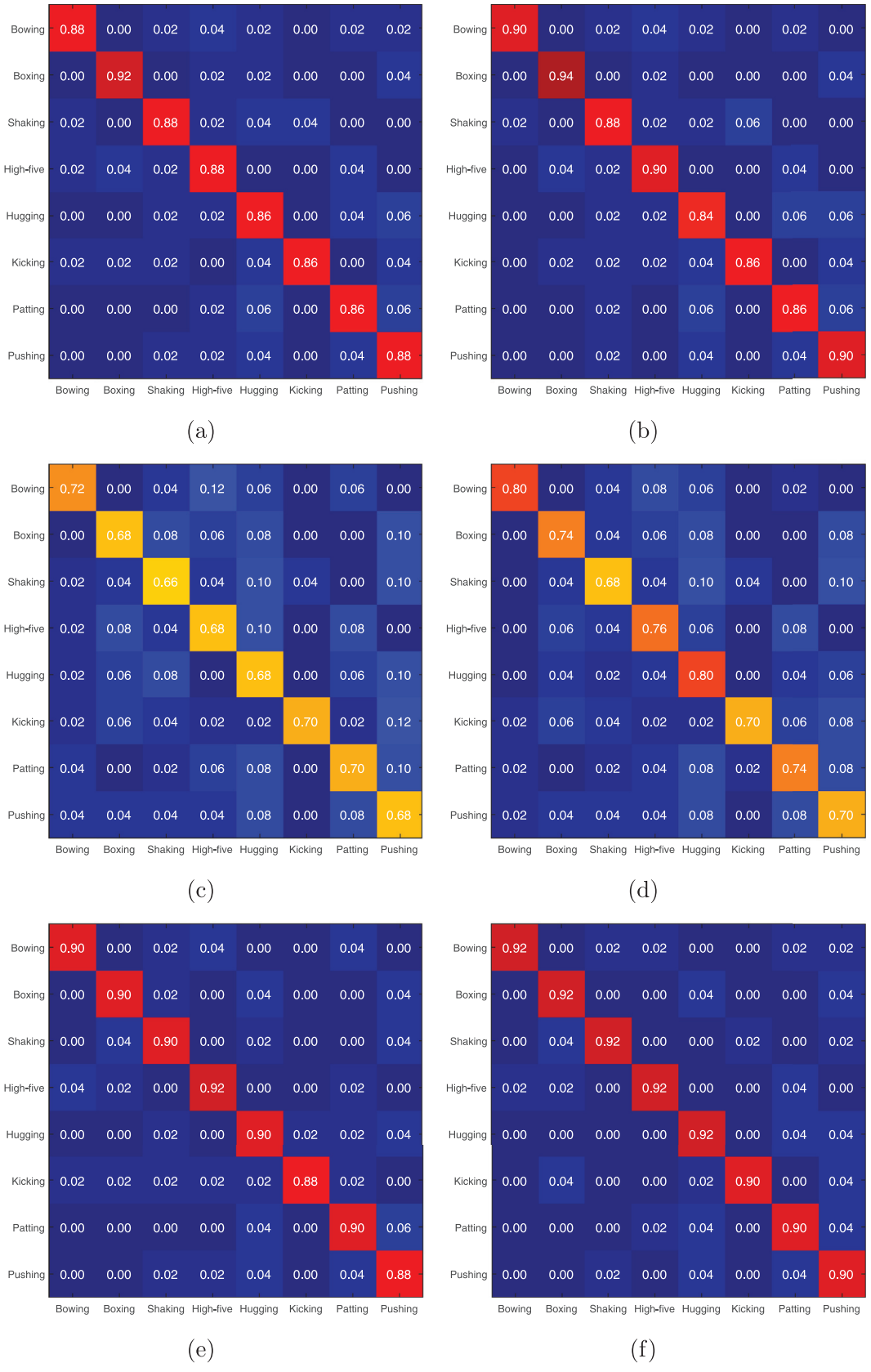

Fig. 6. Confusion matrices of the SVM classifier on the BIT-Interaction data set using 26-part pattern: (a) spatial distance feature, (b) temporal distance feature, (c) spatial angle feature, (d) temporal angle feature, (e) spatial distance-angle feature, and (f) temporal distance-angle feature.

Table 3
Recognition accuracy (%) of the proposed method using different feature types.

BIT-Interaction data set									
Features	Bowing	Boxing	Shaking	High-five	Hugging	Kicking	Patting	Pushing	**Overall**
Spatial distance	88.0	92.0	88.0	88.0	86.0	86.0	86.0	88.0	**87.8**
Temporal distance	90.0	94.0	88.0	90.0	84.0	86.0	86.0	90.0	**88.5**
Spatial angle	72.0	68.0	66.0	68.0	68.0	70.0	70.0	68.0	**68.8**
Temporal angle	80.0	74.0	68.0	76.0	80.0	70.0	74.0	70.0	**74.0**
Spatio-temporal distance	90.0	94.0	90.0	92.0	88.0	88.0	86.0	92.0	**90.0**
Spatio-temporal angle	82.0	74.0	70.0	74.0	78.0	74.0	78.0	76.0	**75.8**
Spatial distance-angle	90.0	90.0	90.0	92.0	90.0	88.0	90.0	88.0	**89.8**
Temporal distance-angle	92.0	92.0	92.0	92.0	92.0	90.0	90.0	90.0	**91.2**
UT-interaction data set									
Features		Shaking	Hugging	Kicking	Pointing	Punching	Pushing	**Overall**	
Spatial distance		80.0	80.0	90.0	90.0	70.0	80.0	**81.7**	
Temporal distance		80.0	90.0	90.0	90.0	80.0	80.0	**85.0**	
Spatial angle		60.0	60.0	70.0	70.0	60.0	60.0	**63.3**	
Temporal angle		70.0	60.0	70.0	70.0	70.0	70.0	**68.3**	
Spatio-temporal distance		90.0	90.0	90.0	100.0	80.0	80.0	**88.3**	
Spatio-temporal angle		70.0	80.0	70.0	80.0	80.0	70.0	**75.0**	
Spatial distance-angle		90.0	90.0	90.0	90.0	80.0	90.0	**88.3**	
Temporal distance-angle		90.0	90.0	90.0	100.0	90.0	90.0	**91.7**	

angle set, temporal angle set, spatio-temporal distance set, spatio-temporal angle set, spatial distance-angle set, and temporal distance-angle set, that were collected from the intra and inter-object distance and angle features in the spatial and temporal dimension. In Fig. 6a–b representing results on the distance feature sets, the proposed method achieves a greater accuracy with the temporal set over the spatial set in most of activities. Compared with the spatial distance set, the temporal distance set contains more information of pose translation. This strategy is repeated on the angle feature sets (Fig. 6c–d) and the merging feature sets of distance and angle (Fig. 6e–f). The above results indicate that the temporal feature sets hold more relational body-part information about object movements. Furthermore, compared with distance feature, the angle information between joint pairs is less useful (68.8% versus 87.8% for spatial feature sets and 74.0% versus 88.8% for temporal feature sets in overall accuracy) because angle feature is quite fragile to noise from the estimation process. When merging distance and angle features following spatial and temporal dimension, the performance in accuracy is differentially improved (see Table 3). However, it is important to note that highly expensive computation is required for merged feature sets. From Table 3, the spatio-temporal distance feature, the spatial distance-angle feature, and the temporal distance-angle feature sets provide the highest overall accuracy results among eight feature categories. According confusion matrices, Hugging and Patting are the most confused together. Hugging and Patting are mostly confused with Pushing (4–6%), and on the contrary. Moreover, Pushing is mostly misunderstood with Hugging and Patting. These activities get some challenges in the pose tracking and locating due to the body-part overlapping. The confusion matrices and the overall accuracy results for the UT-Interaction data set are shown in Fig. 7 and Table 3, respectively. The angle feature is not compatible for this data set with low accuracy (less than 64% and 69% for spatial and temporal angle set, respectively), even if the temporal angle feature category is used. Combining distance and angle features to merging sets in the spatial and temporal dimension sometimes does not bring accuracy improvement at all, for instance, the recognition results of Kicking and Pointing using spatial distance feature in Fig. 7a are equal to results of spatial distance-angle feature in Fig. 7e. In the best performance case using the temporal distance-angle feature set, Pointing is recognized correctly with 100% in accuracy. Punching and Pushing are confused each other due to some resemblances of interactive poselets in the beginning and ending period of activities.

In the last experiment, the proposed method is validated using the merged feature set (see Table 1) that contained information of joint distance and angle features extracted in the spatio-temporal dimension. Two confusion matrices of the SVM classifier corresponding to two interaction data sets are presented in Fig. 8. Compared with feature sets in the second experiment, the proposed method provides higher recognition rates with merged feature sets on the BIT-Interaction dataset (Bowing, Boxing, Hugging, Kicking, Patting, and Pushing) and the UT-Interaction data set (Kicking and Pointing). Nevertheless, the confusions are still occurred with activities involving occlusions, such as Hugging, Patting, Punching, and Pushing. In this experiment, the authors further compare the proposed PAM-based method with the LDA-based approach. Although LDA is constructed on the DAGs structure with Dirichlet distribution, it is only capable to capture the correlation among the features (as codewords) to support directly to the high level information (as activities) without intermediate knowledge from interactive poselets. According to the benefits from capturing correlations among relational features, as well as among interactive poselets and activities, PAM outperforms LDA, greater than 10% and 13% of overall accuracy on two test data sets. Moreover, the authors do an accuracy competition between the proposed method with existing interaction recognition methods, concretely, Lan et al. [26], Ryoo et al. [39], Yu et al. [59], Ryoo et al. [38], Kong et al. [23], and Kong et al. [24] on the same data sets. The recognition accuracy results are presented in Table 4 for the BIT-Interaction data set and the

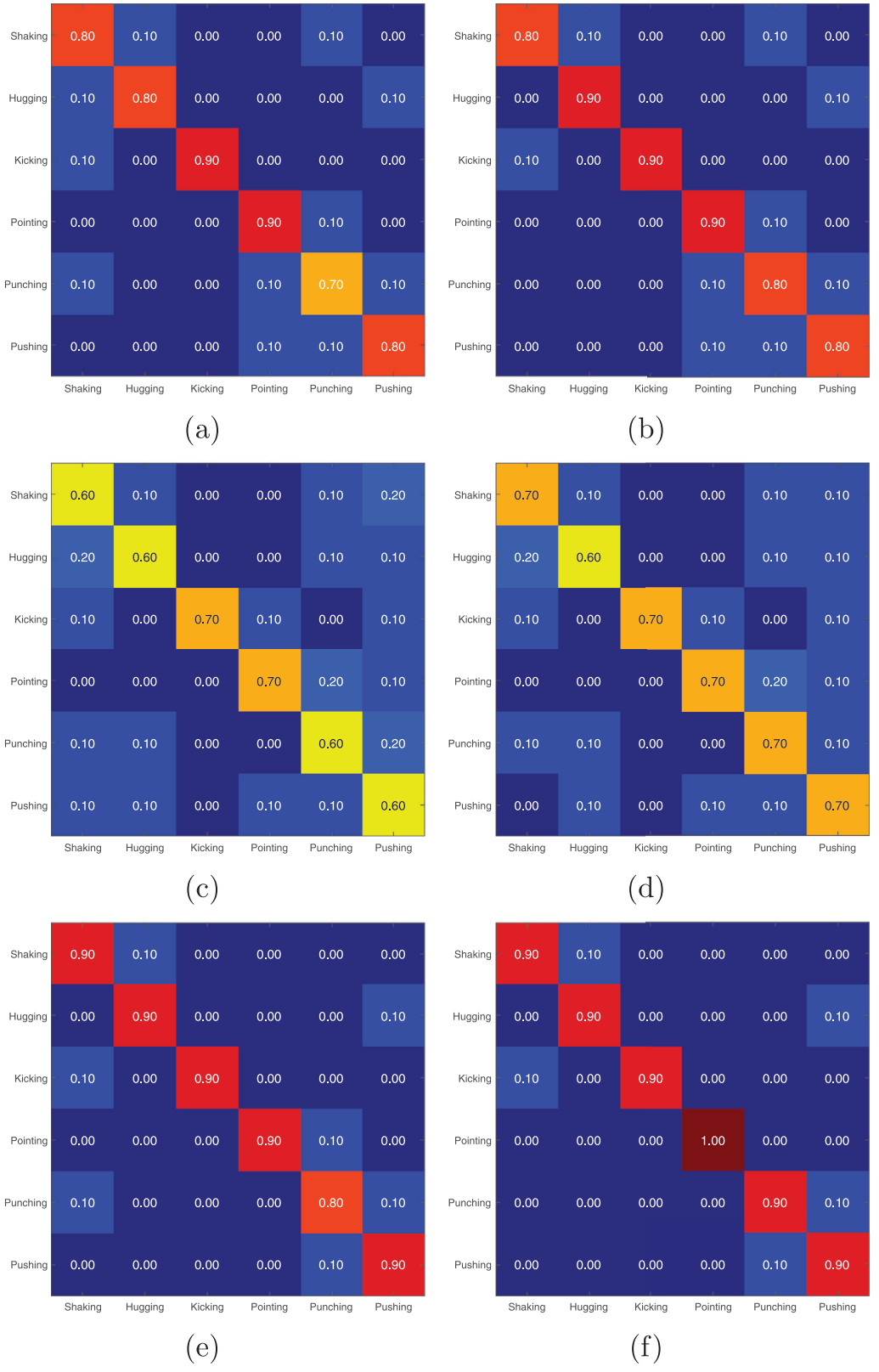

Fig. 7. Confusion matrices of the SVM classifier on the UT-Interaction data set using 26-part pattern: (a) spatial distance feature, (b) temporal distance feature, (c) spatial angle feature, (d) temporal angle feature, (e) spatial distance-angle feature, and (f) temporal distance-angle feature.

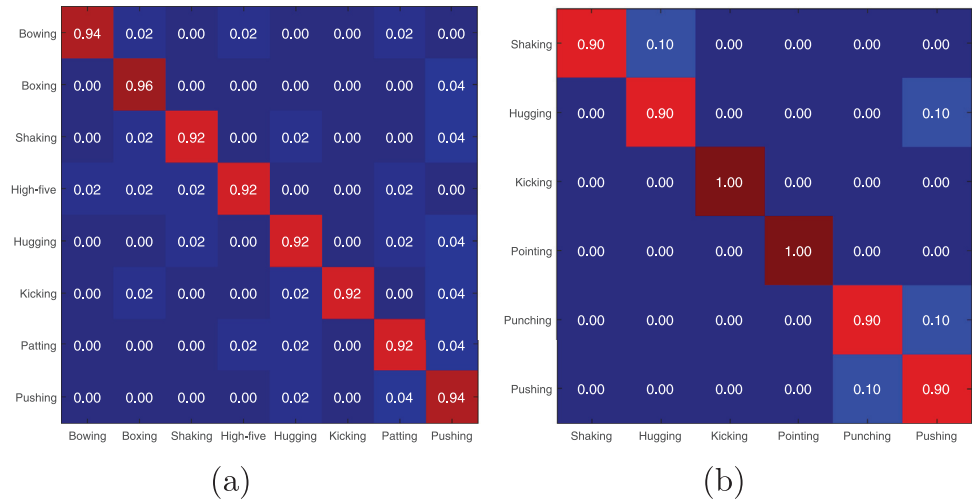

Fig. 8. Confusion matrices of the SVM classifier using 26-part pattern with merged feature category: (a) BIT-Interaction data set, (b) UT-Interaction data set.

Table 4
Comparing recognition accuracy (%) of the proposed method with existing methods.

BIT-interaction data set									
Features	Bowing	Boxing	Shaking	High-five	Hugging	Kicking	Patting	Pushing	Overall
LDA [4]	84.0	88.0	84.0	88.0	84.0	82.0	76.0	78.0	**83.0**
Lan et al. [26]	82.0	76.0	80.0	88.0	88.0	82.0	82.0	80.0	**82.3**
Yu et al. [59]	86.0	84.0	80.0	84.0	82.0	86.0	84.0	80.0	**83.3**
Ryoo et al. [38]	88.0	88.0	80.0	88.0	84.0	88.0	80.0	76.0	**84.0**
Kong et al. [23]	82.0	80.0	82.0	94.0	94.0	80.0	82.0	88.0	**85.3**
Kong et al. [24]	94.0	88.0	94.0	94.0	94.0	88.0	88.0	88.0	**91.0**
Proposed method	94.0	96.0	92.0	92.0	92.0	92.0	92.0	94.0	**93.0**

UT-interaction data set							
Features	Shaking	Hugging	Kicking	Pointing	Punching	Pushing	**Overall**
LDA [4]	80.0	70.0	90.0	90.0	80.0	70.0	**80.0**
Lan et al. [26]	80.0	80.0	100.0	80.0	70.0	70.0	**80.0**
Yu et al. [59]	100.0	80.0	70.0	100.0	80.0	70.0	**83.3**
Ryoo et al. [38]	80.0	90.0	90.0	90.0	80.0	80.0	**85.0**
Kong et al. [23]	80.0	80.0	100.0	90.0	90.0	90.0	**88.3**
Kong et al. [24]	100.0	90.0	100.0	80.0	90.0	90.0	**91.7**
Proposed method	90.0	90.0	100.0	100.0	90.0	90.0	**93.3**

UT-Interaction data set. According to the experimental outcomes, the proposed method outperforms the others in most of testing activities. The activity co-occurrence based method, proposed by Lan et al. [26], combines the adaptive structure and the HOG-based action context descriptor to model the person-person interaction. Due to spatial relation exploration, Lan's model is restricted to deeply understand more complex interactive activities which are generally required more temporal information. Ryoo et al. [38] introduce a novel methodology for activity prediction and recognition based on a dynamic bag-of-words. Although Ryoo's method is capable to fairly handle noise, it is inhibited by overlapping interactions, such as Patting and Pushing because of outliers from the spatio-temporal feature extractor [12]. Building on the work of Ryoo et al. [39], Yu et al. [59] propose Pyramid Spatio–Temporal Relationship Match (PSRM) to combine with Semantic Texton Forest (STFs) to upgrade recognition performance. Video-FAST descriptors provide good performance in processing speed, however, they are corruptible in practical environments containing more dynamic motions. This drawback explains for quite poor accuracy of Yu's method at Pushing, Shaking, and Hugging. Two approaches proposed by Kong et al. [23,24] significantly exceed previous works. According to high-level descriptors, called interactive phrases, Kong et al. formulate binary semantic relationships between interacting people [23]. Concretely, each interactive phrase, detected by an attribute model, is associated with only one attribute belonging to corresponding interactive person to describe motion relationships. Understanding co-occurrence relationships between pairs of interactive phrases therefore addresses motion ambiguity and partial occlusion. In [24], Kong et al. improve recognition performance by a data-driven attribute model and a new learning formulation. The extended version brought some improvements in classification accuracy when compared with original [23], for instance, 91.0% versus 85.3% for the BIT-Interaction data set and 91.7% versus 88.3% for UT-Interaction data set in overall accuracy. Kong's improvement approach recognizes more accurate than the proposed method at High-five, Shaking, and Hugging activities.

Table 5
Processing time result (Mins) of the proposed method.

Step	BIT-interaction (~28,000 frames)	UT-interaction (~10,800 frames)
Pose estimation	20	32
Feature extraction and Codebook construction	58	35
Activity modeling and Classification	212	168
Total	~ 290	~ 235

Learning data-driven phrases using the information bottleneck technique [44] is able to extract discriminative phrases for differentiating interactive activities. In additions, semantic descriptors are able to handle the motion ambiguity and partial occlusion in the interactions while the proposed method is fairly depended on the articulation estimation outcomes. However, the limitation is that the method did not consider dependencies of phrases and attributes in the temporal dimension to lead to some misperception at Pushing, Patting, and Pointing. Different from existing approaches, the proposed method calculates joint distance and angle features from detected joint coordinates using an effective articulated-body estimation to describe intra and inter-person relation in spatio-temporal dimension. The PAM-based hierarchical topic model provides full and flexible correlations of feature-poselet-interaction to maximize explicitness between activities through interactive poselets. Each step in the method is processed separably, the processing time is therefore measured individually and then accumulated for the total. The processing time results are detail listed in Table 5. It is necessary to note that the 26-part pattern and the merged feature set are installed as the default setting for the timing experiment. The processing time for pose estimation depends on the number of frames and the frame resolution. The Gibbs sampling used in the 4-Level PAM for activity modeling is run with 1000 burn-in iterations and its time varies on the number of activities and number of interactive poselets.

5. Conclusion

We proposed a four-level topic model, developed from Pachinko Allocation Model, for the interactive activity recognition, in which the relationships between the relation features and the interactions are fully described through the interactive poselets. In our approach, the intra and inter-person joint features of distance and angle are calculated in the spatio-temporal dimension from the pose estimation outcome. The poselet layer is composed by two types of codeword, d-word and a-word corresponding to joint distance and angle feature, to differentiate the complex interactions. Compared with the 14-part pattern, the proposed method achieves the better accuracy with the 26-part pattern used for pose estimation. Among testing feature categories, the merged feature set is reported as the best results for two benchmark interaction datasets. Moreover, we compare our approach with the standard LDA model used for topic modeling and the state-of-the-art approaches to demonstrate remarkable efficiency in the challenge of interaction recognition.

Because PAM is originally built as a parameter model, the numbers of super topic and sub-topic have to be predefined in advance. The hierarchical Dirichlet process (HDP) [46] can be used as a nonparametric prior for learning the number of topic. Another problem is the high feature dimension that comes from the feature type and the body pattern used in the joint estimation Some advance feature selection algorithms and dimensional reduction techniques are capable to apply for this task without degradation of recognition accuracy.

References

[1] R. Alazrai, Y. Mowafi, C.G. Lee, Anatomical-plane-based representation for humanhuman interactions analysis, Pattern Recognit. 48 (8) (2015) 2346–2363.
[2] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1014–1021.
[3] F.I. Bashir, A.A. Khokhar, D. Schonfeld, Object trajectory-based activity classification and recognition using hidden markov models, IEEE Trans. Image Process. 16 (7) (2007) 1912–1919.
[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[5] X. Cao, Z. Liu, Type-2 fuzzy topic models for human action recognition, Fuzzy Syst. IEEE Trans. PP (99) (2014). 1–1
[6] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27:1–27:27.
[7] Y. Chen, Z. Li, X. Guo, Y. Zhao, A. Cai, A spatio-temporal interest point detector based on vorticity for action recognition, in: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, 2013, pp. 1–6.
[8] Z. Cheng, L. Qin, Q. Huang, S. Yan, Q. Tian, Recognizing human group action by layered model with multiple cues, Neurocomputing 136 (2014) 124–135.
[9] S. Cho, H. Byun, A space-time graph optimization approach based on maximum cliques for action detection, IEEE Trans. Circuits Syst. Video Technol. 26 (4) (2016) 661–672.
[10] W. Choi, S. Savarese, Understanding collective activitiesof people from videos, Pattern Anal. Mach. Intell. IEEE Trans. 36 (6) (2014) 1242–1257.
[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 886–893.
[12] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, 2005, pp. 65–72.
[13] M. Eichner, V. Ferrari, Better appearance models for pictorial structures, in: British Machine Vision Conference, 2009.
[14] A. Eweiwi, M.S. Cheema, C. Bauckhage, Action recognition in still images by learning spatial interest regions from videos, Pattern Recognit. Lett. 51 (2015) 8–15.

[15] B. Fei, J. Liu, Binary tree of svm: a new fast multiclass training and classification algorithm, Neural Netw. IEEE Trans. 17 (3) (2006) 696–704.
[16] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (1) (2005) 55–79.
[17] M.A. Fischler, R. Elschlager, The representation and matching of pictorial structures, Comput. IEEE Trans. C-22 (1) (1973) 67–92.
[18] U. Gaur, Y. Zhu, B. Song, 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, 2011, pp. 2595–2602.
[19] I. Gori, J.K. Aggarwal, L. Matthies, M.S. Ryoo, Multitype activity recognition in robot-centric scenarios, IEEE Robot. Autom. Lett. 1 (1) (2016) 593–600.
[20] S. Jones, L. Shao, Content-based retrieval of human actions from realistic video databases, Inf. Sci. 236 (2013) 56–65.
[21] Y. Kong, Y. Fu, Close human interaction recognition using patch-aware models, IEEE Trans. Image Process. 25 (1) (2016) 167–178.
[22] Y. Kong, Y. Jia, A hierarchical model for human interaction recognition, in: Multimedia and Expo (ICME), 2012 IEEE International Conference on, 2012, pp. 1–6.
[23] Y. Kong, Y. Jia, Y. Fu, Learning human interaction by interactive phrases, in: Computer Vision ÈóÃÇÃô ECCV 2012, in: Lecture Notes in Computer Science, volume7572, 2012, pp. 300–313.
[24] Y. Kong, Y. Jia, Y. Fu, Interactive phrases: Semantic descriptionsfor human interaction recognition, Pattern Anal. Mach. Intel. IEEE Trans. 36 (9) (2014) 1775–1788.
[25] Y. Kong, W. Liang, Z. Dong, Y. Jia, Recognising human interaction from videos by a discriminative model, Comput. Vis. IET 8 (4) (2014) 277–286.
[26] T. Lan, Y. Wang, W. Yang, S. Robinovitch, G. Mori, Discriminative latent models for recognizing contextual group activities, Pattern Anal. Mach. Intel. IEEE Trans. 34 (8) (2012) 1549–1562.
[27] W. Li, A. Mccallum, Pachinko allocation: Scalable mixture models of topic correlations, of Machine Learning Research. Submitted, 2008.
[28] A. Liu, Human action recognition with structured discriminative random fields, Electron. Lett. 47 (11) (2011) 651–653.
[29] L. Liu, L. Shao, X. Zhen, X. Li, Learning discriminative key poses for action recognition, Cybern. IEEE Trans. 43 (6) (2013) 1860–1870.
[30] L. Liu, L. Shao, F. Zheng, X. Li, Realistic action recognition via sparsely-constructed gaussian processes, Pattern Recognit. 47 (12) (2014) 3819–3827.
[31] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and viterbi path searching, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1–8.
[32] L. Meng, L. Qing, P. Yang, J. Miao, X. Chen, D. Metaxas, Activity recognition based on semantic spatial relation, in: Pattern Recognition (ICPR), 2012 21st International Conference on, 2012, pp. 609–612.
[33] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 104–111.
[34] B.T. Morris, M.M. Trivedi, Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach, IEEE Trans. Pattern Anal. Mach. Intel. 33 (11) (2011) 2287–2301.
[35] D. Ramanan, Learning to parse images of articulated bodies, in: Proceedings of Advances in Neural Information Processing System, 2007.
[36] M. Ramanathan, W.-Y. Yau, E.K. Teoh, Human action recognition with video data: Research and evaluation challenges, Human Mach. Syst. IEEE Trans. 44 (5) (2014) 650–663.
[37] M. Rodriguez, C. Orrite, C. Medrano, D. Makris, A time flexible kernel framework for video-based activity recognition, Image Vis. Comput. 48–49 (2016) 26–36.
[38] M. Ryoo, Human activity prediction: Early recognition of ongoing activities from streaming videos, in: Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 1036–1043.
[39] M. Ryoo, J. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 1593–1600.
[40] S. Samanta, B. Chanda, Space-time facet model for human activity classification, Multimed. IEEE Trans. 16 (6) (2014) 1525–1535.
[41] B. Sapp, C. Jordan, B. Taskar, Adaptive pose priors for pictorial structures, in: Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on, 2010, pp. 422–429.
[42] B. Sapp, A. Toshev, B. Taskar, Cascaded models for articulated pose estimation, in: Proceedings of the 11th European Conference on Computer Vision: Part II, in: ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 406–420.
[43] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia, in: MULTIMEDIA '07, 2007, pp. 357–360.
[44] N. Slonim, N. Tishby, Document clustering using word clusters via the information bottleneck method, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '00, ACM, 2000, pp. 208–215.
[45] L. Sun, H. Ai, S. Lao, Localizing activity groups in videos, Comput. Vis. Image Understand. 144 (2016) 144–154.
[46] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical dirichlet processes, J. Am. Stat. Assoc. 101 (2004).
[47] J. Tian, L. Li, 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, 2014, pp. 1–8.
[48] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, C. Sun, Action recognition using nonnegative action component representation and sparse basis selection, Image Process. IEEE Trans. 23 (2) (2014) 570–581.
[49] L. Wang, L. Cheng, T.H. Thi, J. Zhang, Human action recognition from boosted pose estimation, in: Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on, 2010, pp. 308–313.
[50] Y. Wang, G. Mori, Human action recognition by semilatent topic models, Pattern Anal. Mach. Intel. IEEE Trans. 31 (10) (2009) 1762–1774.
[51] E.-J. Weng, L.-C. Fu, On-line human action recognition by combining joint tracking and key pose recognition, in: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, 2012, pp. 4112–4117.
[52] X. Wu, D. Xu, L. Duan, J. Luo, Y. Jia, Action recognition using multilevel features and latent structural svm, Circuits Syst. Video Technol. IEEE Trans. 23 (8) (2013) 1422–1431.
[53] Q. Xiao, J. Cheng, Human action recognition using topic model, in: Information Science and Technology (ICIST), 2014 4th IEEE International Conference on, 2014, pp. 694–697.
[54] J. xin Cai, X. Tang, G. Feng, Learning pose dictionary for human action recognition, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 381–386.
[55] S. Yang, C. Yuan, W. Hu, X. Ding, A hierarchical model based on latent dirichlet allocation for action recognition, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 2613–2618.
[56] W. Yang, Y. Wang, G. Mori, Recognizing human actions from still images with latent poses, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 2030–2037.
[57] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, Pattern Anal. Mach. Intel. IEEE Trans. 35 (12) (2013) 2878–2890.
[58] B. Yao, 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, 2010, pp. 17–24.
[59] T.-H. Yu, T.-K. Kim, R. Cipolla, Real-time action recognition by spatiotemporal semantic and structural forest, in: Proceedings of the British Machine Vision Conference, 2010, pp. 52.1–52.12.
[60] S. Zhang, H. Yao, X. Sun, K. Wang, J. Zhang, X. Lu, Y. Zhang, Action recognition based on overcomplete independent components analysis, Inf. Sci. 281 (2014) 635–647.
[61] Y. Zheng, Y.-J. Zhang, X. Li, B.-D. Liu, Action recognition in still images using a combination of human pose and context information, in: Image Processing (ICIP), 2012 19th IEEE International Conference on, 2012, pp. 785–788.
[62] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: A literature review, Pattern Recognit. 48 (8) (2015) 2329–2345.