An Interactive Activity Recognition Approach Using Articulated-Body Estimation and Pose-Based Features

Thuong Le-Tien
Dept. of Electrical and Electronics Engineering, University of Technology
Hochiminh City, Vietnam
thuongle@hcmut.edu.vn

Thien Huynh-The
Dept. of Computer Engineering, Kyung Hee University
Yongin-si, Gyeonggi-do, Korea
thienht@oslab.khu.ac.kr

Sungyoung Lee
Dept. of Computer Engineering, Kyung Hee University
Yongin-si, Gyeonggi-do, Korea
sylee@oslab.khu.ac.kr

Yongik Yoon
Dept. of Multimedia Science, Sookmyung Women's University
Youngsan-gu, Seoul, Korea
yiyoon@sookmyung.ac.kr

ABSTRACT

In this paper, we go beyond the problem of recognizing human interactions using videos collected from CCTV-based surveillance systems. We propose an approach that permits to deeply describe common person-person activities in the daily life based on the human poses. The joint coordinates of detected human objects are first located by an impressive articulated-body estimation algorithm using the tree graphical structure technique. The relational features consisting of the intra and inter-person feature describing the joint distance and angle information are used for describing the relationships between body components of the individual persons and the interaction of two participants. Moreover, the interaction is also considered in the spatio-temporal dimension in order to upgrade the discrimination among complex activities having much homothetic representation. We validate our interaction recognition method on two practical datasets, the BIT-Interaction dataset and the UT-Interaction dataset, using the multi-class Support Vector Machine technique. The experimental results demonstrate that the proposed approach using pose-body features outperforms recent interaction recognition approaches in the term of classification accuracy.

Categories and Subject Descriptors

Computing methodologies [**Computer Vision**]: Activity recognition and understanding

Keywords

Body Pose Feature, Articulated-body Estimation, Interactive Activity Recognition.

IMCOM '16, January 04-06, 2016, Danang, Viet Nam

DOI: http://dx.doi.org/10.1145/2857546.2857649

1. INTRODUCTION

In recent years, human activity recognition has been an interesting research field in computer vision and artificial intelligence because of wide range of potential applications in indoor-outdoor surveillance and human-computer interaction [1]. Although received more attentions from community, proposing an efficient method for recognizing activities in the real environment still remains a challenging problem due to the large variations of human appearance and other issues, such as mutual occlusion and object interaction. Most existing approaches exploit the low-level features as local spatio-temporal features instead of the human body skeleton for representing observed activities due to the restriction of pose estimation performance, however, some recently impressive researches of human pose computation bring an opportunity for improving recognition accuracy rate.

In computer vision field, human activities can be categorized into the single action and group action. Many researches were introduced for recognizing activities of one actor as walking, jogging, running, hand waving [21], besides the daily life activities in the indoor environment as eating, drinking, typing, and answering phone [14]. Group action, generally performed by visual separable people with complicated interactions, such as walking together, approaching, and gathering, has been investigated using human-based features and movement information for detection and recognition [2, 3]. Few works faces with complex interactive activities, such as hand shaking, hugging, punching, and patting [9], in which object relations should be described for activity representation through discriminative and robust features.

In this work, we present a method for human interaction recognition using the body-pose features extracted in the spatio-temporal dimension. Locating human articulation is performed by an impressive algorithm of pose estimation [22] on detected human objects in a scene. For representing interactions, the spatio-temporal relation features, calculated from the datatset of articulated-pose coordinate, are

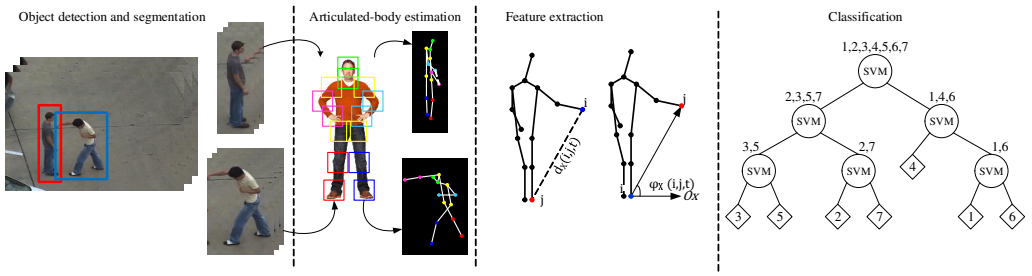


Figure 1: The work flow of a proposed interaction recognition method using spatio-temporal relation features and topic model. The joint coordinates were achieved by using Yang's estimation algorithm on each detected human object. The spatio-temporal relation feature consisting of joint distance and angle between pairs of joints was then extracted. Finally, interactive activities were classified by a Multi-class SVM.

proposed to use, which include the intra and inter-person features of joint distance and angle metric. These features describe the relationships between body components of single persons and interacting participants. Capturing object relationships between the current frame and the previous frame provides more useful knowledge for distinguishing homothetic activities in visualization. Support Vector Machine (SVM) method is finally applied for solving the multi-class classification problem.

2. RELATED WORK

2.1 Human Pose Estimation

Human pose estimation has an important role in the human activity recognition, in which the articulated-pose coordinates or the body part areas are provided corresponding to each human object in the still images. One of most used techniques is the spatial structure coding, often described by the probabilistic graphical model. Although structural-based graphical model allows for inferring efficient body parts, it sometimes fails in locating body parts. Based on the pictorial structure model introduced by Fischler et al. [8], Huttenlocher et al. [7] presented an object as a collection of arranged parts in a deformable configuration. Eichner et al. [5] focused on improving the estimation accuracy for unusual poses by exploiting the latent relationships between the appearance of different body parts from annotated images. Body parts were plugged into any pictorial structure engines using learned appearance models. A shape-based kernel for upper-body pose similarity and a leave-one-out loss function were further developed for learning task. In recent years, a general and flexible mixture model introduced by Yang et al. [22] using the standard pictorial structure model captured the spatial relations between part positions and the co-occurrence relations between part mixtures. Most RGB-based pose estimation approaches have been restricted in component identifying and locating due to variations of appearance, mutual occlusion, and among others.

2.2 Features For Activity Representation

Based on the feature type criteria, activity recognition approaches are categorized into two groups: local spatio-temporal interest point (STIP) [4, 19, 18, 17] and body-pose feature [21, 24, 13, 20]. Local feature-based approaches have shown practical limitations of non-robustness in dynamic scenes, low recognition accuracy with complex activities, and poor relation description in interactions. Wu et al. [19] took advantage of the interest point detector, proposed by Dollar et al.[4], including 2D Gaussian filter and 1D Gabor filter to produce a high response at each detected point. Some well-known feature descriptors were usually applied for feature extraction, such as Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and Scale-Invariant Feature Transform (SIFT), Features From Accelerated Segment Test (FAST), and Motion-Constrained SIFT (MoSIFT). A novel feature based on 3D Haar wavelet transform was suggested by Samanta et al. [17] for space-time interest points to restrict outliers from the feature extraction process. Compared with STIP-based approaches, pose-based methods proved the advantage and flexibility in recognizing complex actions and interactions, however, the recognition accuracy mostly depends on the pose estimation results. Different from usual approaches in separating the pose estimation and the action recognition stage, Yang et al. [21] trained an integrated fashion system that jointly considered poses and actions to directly obtain the pose information. Poselet Activation Vector in [24] consisted of pose information was combined with contextual information for learning the poselet-based action classifier. Extensive pyramidal feature (EPF), constructed from three components of the Gabor filter, Gaussian pyramid, and wavelet transform, was proposed for describing poses by Liu et al. [13]. A pose dictionary established by shape of contour points from the human silhouette was formulated by Cai et al. [20] for the task of activity recognition.

3. METHODOLOGY

The proposed interaction recognition method consists of the articulated-body estimation, spatio-temporal relation feature extraction, and classification module as Figure 1.

3.1 Articulated-body estimation

In this work, an impressive articulated-body estimation algorithm, introduced by Yang et al. [22], is used for locating the body joint coordinates of detected human objects. The human key points were modeled into a tree graphical structure, and a score function was the formulated to flexibly detect human and search poses efficiently in images. The full core function is formulated based on a compatibility function and a configuration of part types and positions, in which a compatibility function is a sum of local and pairwise scores. Based on capturing dependence information of local

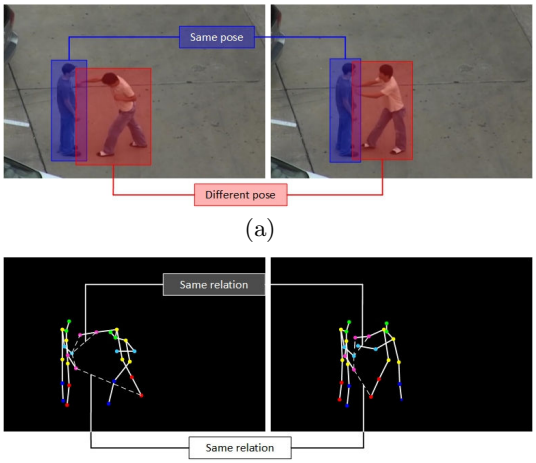


Figure 2: Interactive activities: punching and pushing contain one person with the same pose.

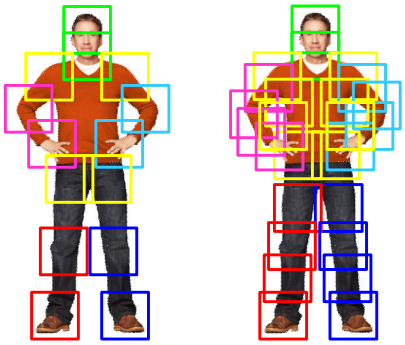


Figure 3: An illustration of articulated-body estimation using the Yang's algorithm: (a) 14-joint pattern, (b) 26-joint pattern.

appearance on the spatial geometry, Yang's model showed the higher speed and accuracy in the estimation task when compared with classic articulation models on the real-life datasets.

In order to boost the performance of articulated joint locating, multiple pose estimators are trained on the testing dataset hereafter to control variance among activities. In the training stage, samples of a particular activity are chosen as positive samples and remainders as negative samples. Therefore each estimator is tried one by one to select the best result with the maximum score in the testing stage. By this strategy, the estimation accuracy is fairly improved. An example for articulation estimation with 14-joint and 26-joint pattern is shown in Figure 3. The dataset of 2D joint coordinates of detected objects in frames is obtained as the output of this stage.

3.2 Body-Pose Feature Extraction

In the interaction, the human poses should be discriminated on different activities. The pose of individual human object are good enough for recognizing actions of a single person, however, for the task of interaction recognition, the relations between two objects have to be exploited due to the pose sharing. In the UT-Interaction dataset [16], Punching and Pushing, which have one human object in the standing pose and another in the acting pose as illustrated in Figure 2, should be represented through the active poses and the object-object relations. Moreover, observing these objects in the temporal dimension gives more expensive information about the pose translation. The spatio-temporal relation features are therefore studied from the pose coordinate dataset. The authors calculate the distance of two joints and angle of joint vector and horizontal axis. Figure 4 describes eight feature types extracted from two objects.

Intra-spatio joint distance: The joint distance feature (see Figure 4a) is defined as the Euclidean distance between a pair of two joints for each human object in a frame; therefore calculated as follows:

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

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

Intra-spatio joint angle: The joint angle feature (see Figure 4b) is defined as the angle between the joint vector $\overrightarrow{p_i p_j}$ and the horizontal axis $\overrightarrow{Ox}$:

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

Inter-spatio joint distance: The inter-spatio joint distance feature (see Figure 4c) is calculated by Eq. 1, where joints belong to different objects. Particularly, it is measured as follows:

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

where $\left\{p_{i,X}^t, p_{j,Y}^t\right\} \in \Re^2$ 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 t^{th} frame.

Inter-spatio joint angle: The inter-spatio joint angle feature (see Figure 4d) is developed from Eq. 2 for two objects:

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

Intra-temporal joint distance: The intra-temporal joint distance (see Figure 4e) represents the Euclidean distance between pair of joints belonging to one human object at the current t^{th} frame and the previous $\left(t-t_0\right)^{th}$ frame:

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

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

Intra-temporal joint angle: The intra-temporal joint angle (see Figure 4f) describes the angle between the joint vector $\overrightarrow{p_i^{t-t_0} p_j^t}$ and the horizontal axis:

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

Inter-temporal joint distance: The inter-temporal joint distance (see Figure 4g) formulates the Euclidean distance between pairs of joints belonging to two different objects at different frames:

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

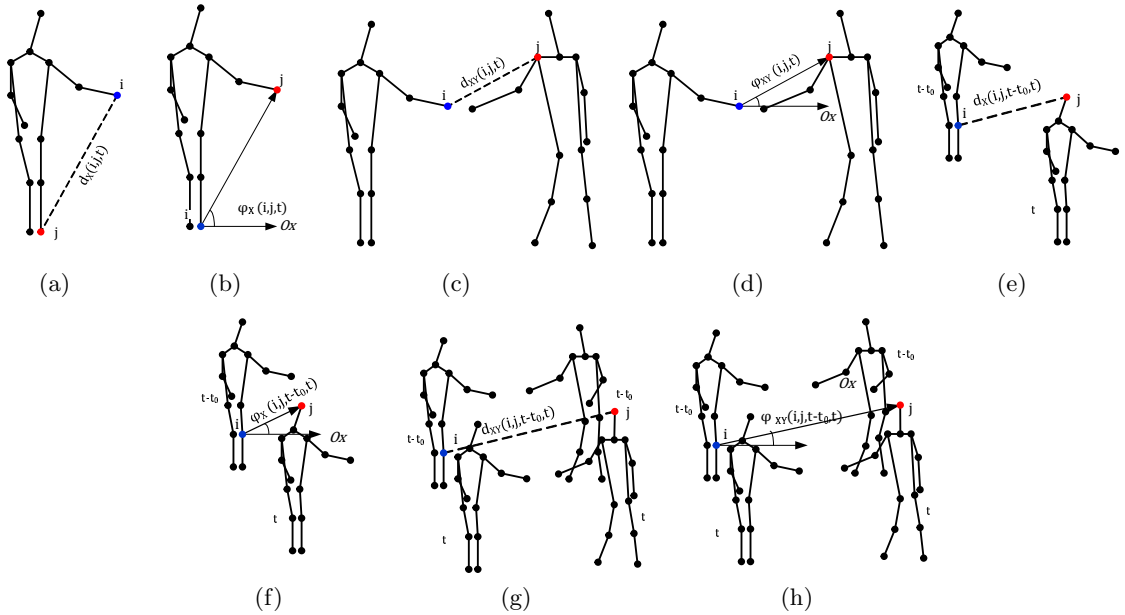


Figure 4: Illustrations of extracted features using the joint coordinate dataset: (a) Intra-spatio joint distance, (b) Intra-spatio joint angle, (c) Inter-spatio joint distance, (d) Inter-spatio joint angle, (e) Intra-temporal joint distance, (f) Intra-temporal joint angle, (g) Inter-temporal joint distance, (h) Inter-temporal joint angle.

where $d_{XY}\left(i,j,t-t_0,t\right)$ 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 $d_{YX}\left(i,j,t-t_0,t\right)$.

Inter-temporal joint angle: Similarly, the inter-temporal joint angle (see Figure 4h) expresses the angle features between two different objects in different frames:

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

Due to the difference in unit, distance and angle features have to be normalized when merging them for recognition. The normalization process is executed as follows:

$$\begin{aligned}\hat{d} &= \left(d-d_{\min}\right)/\left(d_{\max}-d_{\min}\right)\\ \hat{\varphi} &= \varphi/2\pi\end{aligned} \tag{9}$$

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

$$\begin{aligned}
D_X^S &= \left\{\hat{d}\left(i,j,t\right)\middle| i\in X^t, j\in X^t, i\neq j\right\}\\
D_Y^S &= \left\{\hat{d}\left(i,j,t\right)\middle| i\in Y^t, j\in Y^t, i\neq j\right\}\\
D_{XY}^S &= \left\{\hat{d}\left(i,j,t\right)\middle| i\in X^t, j\in Y^t\right\}\\
D_X^T &= \left\{\hat{d}\left(i,j,t-t_0,t\right)\middle| i\in X^{t-t_0}, j\in X^t\right\}\\
D_Y^T &= \left\{\hat{d}\left(i,j,t-t_0,t\right)\middle| i\in Y^{t-t_0}, j\in Y^t\right\}\\
D_{XY}^T &= \left\{\hat{d}\left(i,j,t-t_0,t\right)\middle| i\in X^{t-t_0}, j\in Y^t\right\}\\
D_{YX}^T &= \left\{\hat{d}\left(i,j,t-t_0,t\right)\middle| i\in Y^{t-t_0}, j\in X^t\right\}\\
\Phi_X^S &= \left\{\hat{\varphi}\left(i,j,t\right)\middle| i\in X^t, j\in X^t, i\neq j\right\}\\
\Phi_Y^S &= \left\{\hat{\varphi}\left(i,j,t\right)\middle| i\in Y^t, j\in Y^t, i\neq j\right\}\\
\Phi_{XY}^S &= \left\{\hat{\varphi}\left(i,j,t\right)\middle| i\in X^t, j\in Y^t\right\}\\
\Phi_X^T &= \left\{\hat{\varphi}\left(i,j,t-t_0,t\right)\middle| i\in X^{t-t_0}, j\in X^t\right\}\\
\Phi_Y^T &= \left\{\hat{\varphi}\left(i,j,t-t_0,t\right)\middle| i\in Y^{t-t_0}, j\in Y^t\right\}\\
\Phi_{XY}^T &= \left\{\hat{\varphi}\left(i,j,t-t_0,t\right)\middle| i\in X^{t-t_0}, j\in Y^t\right\}\\
\Phi_{YX}^T &= \left\{\hat{\varphi}\left(i,j,t-t_0,t\right)\middle| i\in Y^{t-t_0}, j\in X^t\right\}
\end{aligned} \tag{10}$$

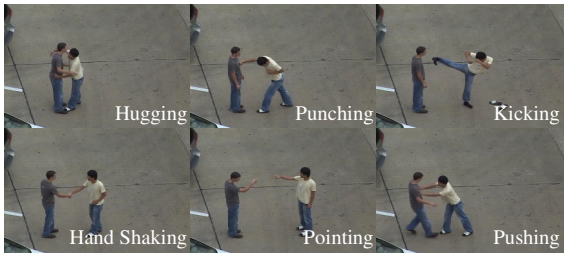
3.3 Classification

To solve the N-class pattern recognition problem, the authors utilize the Binary Tree of SVM [6], 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 should be assigned 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. An essential contribution of the BTS-SVM approach, the multiclass issue, is mapped into binary-tree architectures without performance reduction.

Table 1: Category Of Extracted Features

Feature Category	Term
Spatial Distance	D_X^S, D_Y^S, D_{XY}^S
Temporal Distance	$D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$
Spatial Angle	$\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S$
Temporal Angle	$\Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$
Spatio-Temporal Distance	$D_X^S, D_Y^S, D_{XY}^S, D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$
Spatio-Temporal Angle	$\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$
Spatial Distance-Angle	$D_X^S, D_Y^S, D_{XY}^S, \Phi_X^S, \Phi_Y^S, \Phi_{XY}^S$
Temporal Distance-Angle	$D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$
Merged Feature	$D_X^S, D_Y^S, D_{XY}^S, D_X^T, D_Y^T, D_{XY}^T, D_{YX}^T$ $\Phi_X^S, \Phi_Y^S, \Phi_{XY}^S, \Phi_X^T, \Phi_Y^T, \Phi_{XY}^T, \Phi_{YX}^T$

(a)



(b)

Figure 5: Two Interaction dataset for evaluation: (a) BIT-Interaction dataset and(b) UT-Interaction dataset.

4. EXPERIMENTS AND DISCUSSIONS

4.1 Dataset and Experiment Setup

In this paper, the proposed method is benchmarked on two well-known interaction datasets, BIT-Interaction dataset [11] and UT-Interaction dataset [16] .

BIT-Interaction dataset has eight classes of human interactive activity comprising *Bowing*, *Boxing*, *Hand Shaking*, $High-five$, *Hugging*, *Kicking*, *Patting*, and *Pushing* as shown in Figure 5a, with 50 short videos (2 seconds) per class. 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 dataset consists of six interactions as *Hand Shaking*, *Hugging*, *Kicking*, *Pointing*, *Punching* and *Pushing* as shown in Figure 5b. Each interaction is presented by 10 videos whose lengths are around 1 minute. Totally, there are 60 videos for six classes provided in the dataset.

In two datasets, the individual human objects in each frame are extracted by their bounding boxes supported in dataset owners for articulated-body 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. The authors only consider the 26-joint pattern instead of 14-joint one due to its higher estimation accuracy [22]. All of the experiments are performed using Matlab 2013a for simulation on a desktop PC running Windows 7 Operating System with a 2.67-GHz Intel Core i5 CPU and 4GB of RAM. In the BTS-SVM classifier, the Radial Basic Function (RBF) kernel is used to set up for each node of binary classification. The proposed method and other approaches used for comparison are evaluated using the 10-fold cross-validation.

4.2 Experiment Results and Discussions

In the first experiment, the authors investigate the proposed method on different feature categories based on the 26-joint pattern. The classification results are reported by the confusion matrices in Figure 6 for the BIT-Interaction dataset. Totally, there are eight considered categories: the spatial distance set, temporal distance set, spatial angle set, temporal angle set, spatio-temporal distance set, spatio-temporal angle set, spatial distance-angle set, and temporal distance-angle set, that are collected from the intra and inter-object joint distance and angle in the spatial and temporal dimension. In Figure 6a-b representing results of the distance feature sets, the proposed method achieved a greater accuracy with the temporal set over the spatial set in most of activity classes. Compared with the spatial distance set, the temporal distance set contains more information of pose translation. This strategy is repeated again on the angle feature set as showed in Figure 6c-d. These results indicate that the temporal feature sets contain more relational body-part information of the object movement. Compared with distance feature, the angle information between joint pairs can not deliver a desirable recognition rate (67.75% versus 85.75% for spatial feature sets and 73.25% versus 87.00% for temporal feature sets in an overall accuracy) because angle are more sensitive to noise than distance. When merging distance and angle features following spatial and temporal dimension, the performance in accuracy is improved with different grades (see Figure 6e-h). However, it is important to note that a more expensive computational cost is required for these merged feature sets. The spatio-temporal distance feature, the spatial distance-angle feature, and the temporal distance-angle feature sets have the highest overall accuracy values at 88.5%, 88.5%, and 90.0%, respectively, among eight investigated feature categories. According the confusion matrices, *Hugging* and *Patting* are the most confused activities in all cases. *Hugging* and *Patting* are mostly confused with *Pushing*; and on the contrary, *Pushing* take the most misclassifications with *Hugging* and *Patting*. These activities get some challenges in the pose tracking and locating due to the body-part overlapping. The confusion matrices for the UT-Interaction dataset are shown in Figure 7. The angle features are not compatible for this dataset with low accuracy (less than 61.67% and 65.00% for spatial and temporal angle set, respectively), even if the temporal angle feature category is applied. Combining distance and angle features to merging sets in the spatial and temporal dimension sometimes can not bring the accuracy improvement at all. *Punching* and *Pushing* are confused each other due to some resemblances of interaction in the beginning and ending period of activities.

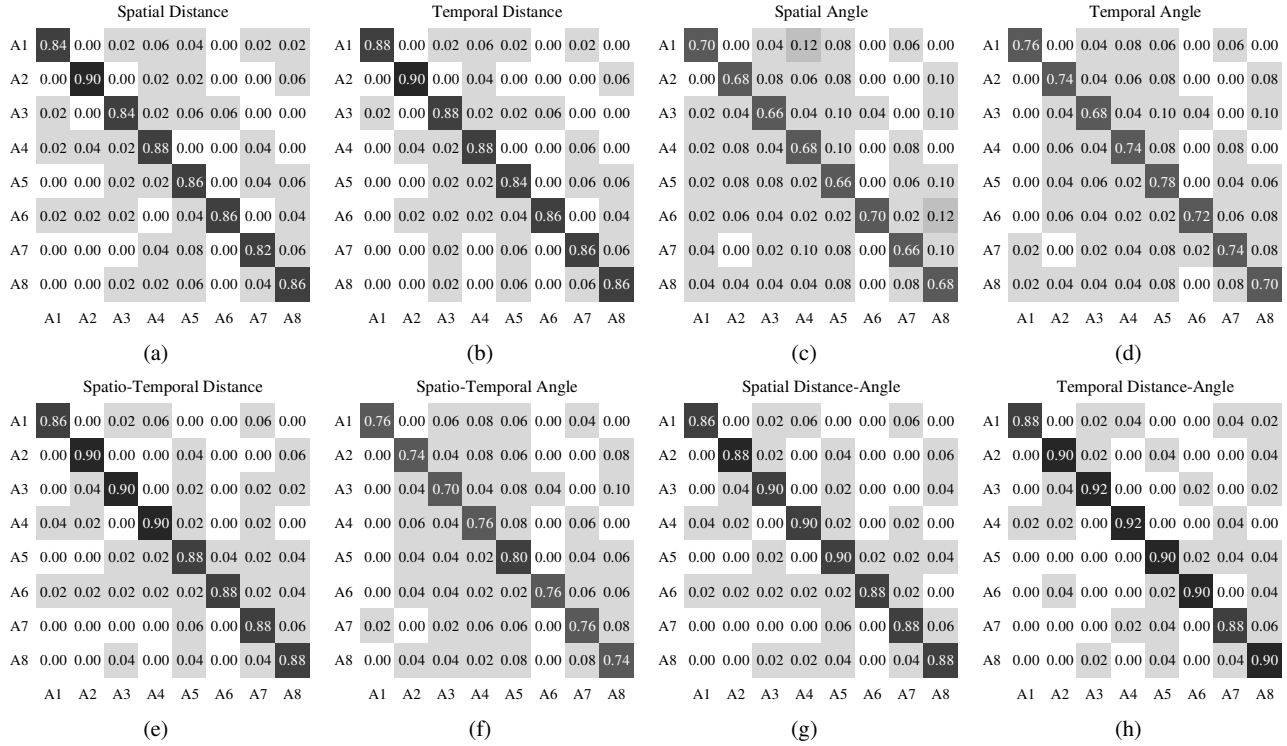


Figure 6: Confusion matrices of the SVM classifier on the BIT-Interaction dataset with various feature categories (A1-A8: bowing, boxing, hand shaking, high-five, hugging, kicking, patting, pushing).

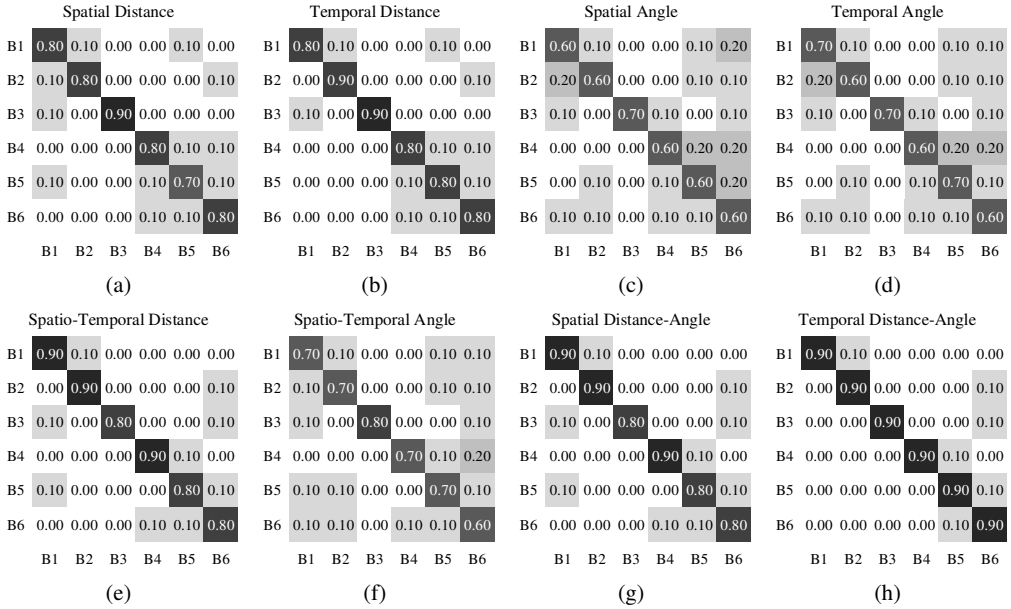


Figure 7: Confusion matrices of the SVM classifier on the UT-Interaction dataset with various feature categories (B1-B6: hand shaking, hugging, kicking, pointing, punching, pushing).

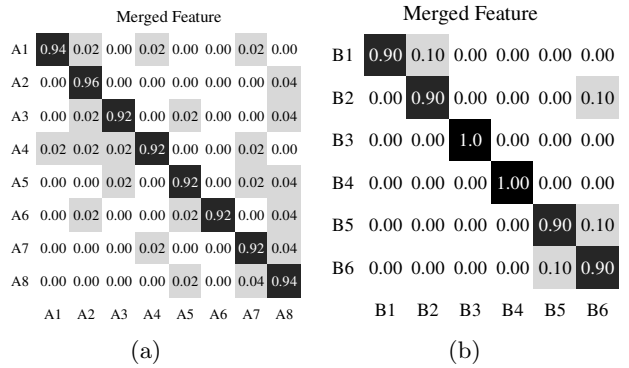


Figure 8: Confusion matrices of the SVM classifier using the merged feature category on: (a) BIT-Interaction dataset, (b) UT-Interaction dataset.

In the second experiment, the proposed method is validated on two interaction datasets using the merged feature set (see Table 1) that contained information of joint distance and angle features extracted in the spatio-temporal dimension. The confusion matrices of the SVM classifier are presented in Figure 8. Compared with investigated feature sets in the first experiment, the proposed method obtains a greater accuracy rate with merged feature sets on the BIT-Interaction and the UT-Interaction dataset at most interactions. The confusions are still occurred in activities involving occlusions, such as *Hugging*, *Patting*, *Punchin*, and *Pushing*. Although acquires an impressive accuracy, utilizing the merged feature set degrades the performance in processing speed.

In this experiment, the authors further make a comparison on classification accuracy between the proposed method with existing impressive interaction recognition methods, concretely, Lan et al. [12], Yu et al. [23], Ryoo et al. [15], and Kong et al. [10] on the same datasets. The results are presented in Table 2 and 3 for the BIT-Interaction and UT-Interaction dataset, respectively. The results prove that the proposed method outperforms the compared methods in most of activities. The action co-occurrence based method, suggested by Lan et al. [12], integrated the adaptive structure and the HOG-based action context descriptor for modeling the person-person interaction. However, Lan's model was restricted in deeply understanding more complex interactive activities which were generally required more temporal information because of only considering spatial object relation instead of temporal relations. Based on the dynamic bag-of-words technique, Ryoo et al. [15] presented an impressive methodology for activity prediction and recognition. Although Ryoo's method was able to quite handle noisy observations, it got the challenge with overlapping interactions, such as *Patting* and *Pushing* because of outliers from the spatio-temporal feature extraction module [4]. Extending the work of Ryoo et al. [16], Yu et al. [23] proposed Pyramid Spatio-Temporal Relationship Match (PSRM) to combine with Semantic Texton Forest (STFs) for accuracy improvement. Utilizing Video-FAST interest points brought a high performance in processing speed, however, they look corruptible in practical environments containing more dynamic motions. This limitation was explained for quite poor accuracy of Yu's method at *Pushing*, *Shaking*, and

Table 2: Method comparison on BIT dataset

Activity	Lan [11]	Yu [22]	Ryoo [14]	Kong [9]	Proposed
Bowing	82	86	88	82	**88**
Boxing	76	84	88	80	**90**
Hand shaking	80	80	80	82	**92**
High-five	88	84	88	**94**	92
Hugging	88	82	84	**94**	92
Kicking	82	86	88	80	**90**
Patting	82	84	80	82	**88**
Pushing	80	80	76	88	**90**
Average	82.3	83.3	84.0	85.3	**89.0**

Table 3: Method comparison on UT dataset

Activity	Lan [11]	Yu [22]	Ryoo [14]	Kong [9]	Proposed
Hand shaking	80	**100**	80	80	90
Hugging	80	80	90	80	**90**
Kicking	100	70	90	**100**	90
Pointing	80	100	90	90	**100**
Punching	70	80	80	90	**90**
Pushing	70	70	80	90	**90**
Average	80.0	83.3	85.0	88.3	**91.6**

Hugging class. The approach proposed by Kong et al. [10] significantly outperformed previous works. By proposing high-level descriptors, called interactive phrases, Kong formulated binary semantic relationships between interacting people with their interaction in [10]. To describe the motion relationships, each interactive phrase, detected by an attribute model, was associated with only one attribute belonging to people in interactions. Therefore, the challenges of motion indistinctness and partial occlusion have been solved based on understanding co-occurrence relationships between pairs of interactive phrases. However, the limitation is that the method did not consider dependencies of phrases and attributes in the temporal dimension to lead to non-perception at *Pushing*, *Patting*, and *Pointing*. Different from compared methods, the proposed approach in this paper calculates joint distance and angle features from data of body joint coordinate using an impressive articulated-body estimation to describe intra and inter-person relation in spatio-temporal dimension.

5. CONCLUSIONS

In this research, we proposed an efficient video-based activity recognition method using the body-pose features. An impressive articulation estimation algorithm is utilized for body joint coordinate extraction. For describing interactive activities between detected human objects, the intra- and inter-person relational features are calculated from coordinate dataset with the joint distance and angle metric. The interactions are considered not only within a same frame but also in two consecutive frames for describing pose translation. The proposed method is investigated and validated on various feature categories that are grouped following the feature types. Temporal feature sets provide a greater recognition accuracy when compared with spatial feature sets; and distance feature is better than angle feature in the most of cases. When merging all features including distance and an-

gle in the spatio-temporal relation, our method reports the best result in classification accuracy. However, the tradeoff is the higher computational cost required for feature extraction and classification process. Advanced feature selection algorithms and dimensional reduction techniques can be applied without accuracy degradation in the future.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazan. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Systems with Applications*, 42(21):7991 – 8005, 2015.

[2] Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian. Recognizing human group action by layered model with multiple cues. *Neurocomputing*, 136:124 – 135, 2014.

[3] W. Choi and S. Savarese. Understanding collective activitiesof people from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1242–1257, June 2014.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceeding of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, October 2005.

[5] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proceedings of The British Machine Vision Conference*, September 2009.

[6] B. Fei and J. Liu. Binary tree of svm: a new fast multiclass training and classification algorithm. *IEEE Transactions on Neural Networks*, 17(3):696–704, May 2006.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, January 2005.

[8] M. A. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, C-22(1):67–92, January 1973.

[9] Y. Kong and Y. Jia. A hierarchical model for human interaction recognition. In *Proceeding of IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2012.

[10] Y. Kong, Y. Jia, and Y. Fu. Learning human interaction by interactive phrases. In *Lecture Notes in Computer Science - European Conference on Computer Vision*, pages 300–313, 2012.

[11] Y. Kong, Y. Jia, and Y. Fu. Interactive phrases: Semantic descriptionsfor human interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1775–1788, September 2014.

[12] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, August 2012.

[13] L. Liu, L. Shao, X. Zhen, and X. Li. Learning discriminative key poses for action recognition. *IEEE Transactions on Cybernetics*, 43(6):1860–1870, December 2013.

[14] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proceeding of IEEE 12th International Conference on Computer Vision*, pages 104–111, September 2009.

[15] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceeding of IEEE International Conference on Computer Vision*, pages 1036–1043, November 2011.

[16] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Proceeding of IEEE 12th International Conference on Computer Vision*, pages 1593–1600, September 2009.

[17] S. Samanta and B. Chanda. Space-time facet model for human activity classification. *IEEE Transactions on Multimedia*, 16(6):1525–1535, October 2014.

[18] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, and C. Sun. Action recognition using nonnegative action component representation and sparse basis selection. *IEEE Transactions on Image Processing*, 23(2):570–581, February 2014.

[19] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia. Action recognition using multilevel features and latent structural svm. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(8):1422–1431, August 2013.

[20] J. xin Cai, X. Tang, and G. Feng. Learning pose dictionary for human action recognition. In *Proceeding of International Conference on Pattern Recognition*, pages 381–386, Aug 2014.

[21] W. Yang, W. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2030–2037, June 2010.

[22] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, December 2013.

[23] T.-H. Yu, T.-K. Kim, and R. Cipolla. Real-time action recognition by spatiotemporal semantic and structural forest. In *Proceedings of the British Machine Vision Conference*, pages 52.1–52.12, 2010.

[24] Y. Zheng, Y.-J. Zhang, X. Li, and B.-D. Liu. Action recognition in still images using a combination of human pose and context information. In *Proceeding of IEEE International Conference on Image Processing*, pages 785–788, September 2012.