# Interactive Activity Recognition Using Articulated-Pose Features on Spatio-Temporal Relation

Thien Huynh-The[1], Dinh-Mao Bui[1], Sungyoung Lee[1], Yongik Yoon[2],

[1]Department of Computer Engineering, Kyung Hee University,
Gyeonggi-do, 446-701, Korea
[2]Department of Multimedia Science, Sookmyung Women's University,
Seoul, 140-172, Korea

{thienht,mao,sylee}@oslab.khu.ac.kr, yiyoon@sookmyung.ac.kr

**Abstract.** A success progress of pose estimation approaches motivates the activity recognition used in CCTV-based surveillance systems. In this paper, a method is proposed for recognizing interactive activities between two human objects. Based on articulated joint coordinates obtained from a pose estimation algorithm, the distance and direction feature are extracted from objects to describe both the spatial and temporal relation. The multiclass Support Vector Machine is finally employed for activity classification task. Compared with existing methods using the public interaction dataset, the proposed method outperforms in overall classification accuracy.

**Keywords:** Interactive Activity Recognition, Articulated-Pose Feature, Spatio-Temporal Relation.

## 1 Introduction

In recent years, human activity recognition has been received more attentions from computer vision and artificial intelligence community due to its important role in applications of video-based surveillance, video annotation, somatic game and human-computer interaction. However, developing an effective system is a challenge due to many issues, such as illumination variation and object occlusion.

Many researches were proposed for activity recognition with one actor, such as running, walking, waving hands [1]. Some methods considered the interaction activities between the actor and object, i.e., the daily life activities as eating, drinking [2] by using the local features. Few works [3] faced with activities involving two or more human objects in interaction likes as hand shaking, hugging, and punching.

One of the most important issues explored in the activity recognition is the feature extraction. Motivated by existing object recognition approaches using local features such as SIFT descriptor [4], which captured the spatial relations between the points of interest. Recently, authors have been attracted by research direction of pose estimation [5], in which the relative locations of body components are identified.

Sparse coding and dictionary learning are two advanced techniques for modeling activities based on extracted features. Unlike the existing dictionary learning approaches utilized the spatio-temporal features, Cai et al. [6] recommended a framework of learning pose dictionary for the human body representation.

In this paper, we focus on the interactive activities between two persons based on the successful outcome of the pose estimation approach [5]. From the articulated-pose coordinates, two kinds of feature, joint distance and joint direction, are extracted to describe the spatial relationship between body components that belong to each object and two objects. Furthermore, the objects are consider not only within a same frame but also in two adjacent frames to enhance the distinctness of benchmarked activities. Finally, a multiclass Support Vector Machine (SVM) classifier is performed for training and classification. In the experiments, we implement the proposed scheme under the different extracted feature categories and compare with the state-of-the-art methods.

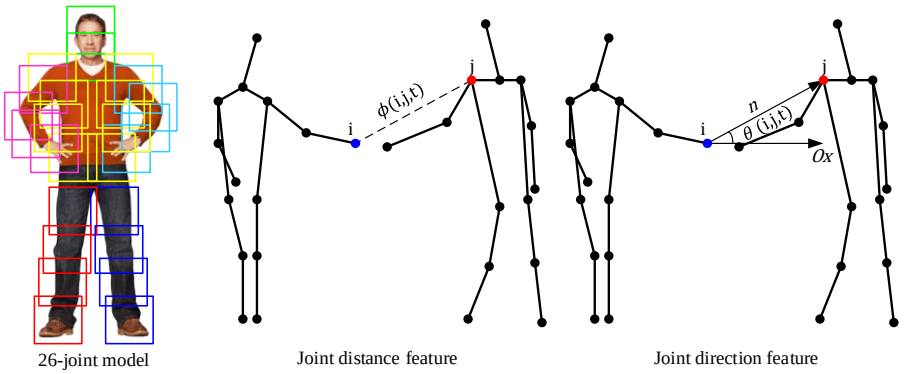


**Fig. 2.** An illustration of: (a) 26 articulated joint model and (b) distance and direction feature.

## 2 Methodology

### 2.1 Joint Detection

One of the most well-known pose estimation methods was introduced by Yang et al. [5], in which, the body key joint were modeled into a tree structure and a score function was identified to dynamically search human object and effectively detect articulation. Based on capturing the dependence of local appearance on the spatial geometry, Yang's model outperformed classic articulated models in speed and accuracy. The impressive results on several real-life datasets were reported in that research to demonstrate the preeminence of this algorithm when compared with existing algorithms in the task of human pose estimation. In this research, we

employed Yang's algorithm for the joint estimation purpose on the interaction dataset [7]. Concretely, the PARSE model with full-body estimation has been used to locate 26 articulated points, shown in Fig. 1, due to the highest accuracy in comparison of 14-point and 51-point model. For enhancing performance of locating articulated points, multiple estimators are trained on the interaction dataset to control huge variance among activities. Each model is tried to detect joints and the best one is selected with maximum score.

### 2.2 Feature Extraction

The input data in this stage contains the coordinates of body joints estimated from Yang's algorithm. Each joint is denoted as $p_i(x, y)$. Two types of feature calculated from the joint coordinates are the distance and the direction, illustrated in Fig. 1.

**Joint distance:** The joint distance feature is defined as the Euclidean distance between all pairs of joints in each person and between two person in a frame and in two adjacent frames. The distance between two joints $i$ of an object $X$ and $j$ of an object $Y$ in the frame $t$ is calculated as follows:

$$\phi^{X,Y}(i,j,t) = \left\| p_{i,t}^{X} - p_{j,t}^{Y} \right\| = \sqrt{\left( x_{p_{i,t}^{X}} - x_{p_{j,t}^{Y}} \right)^2 + \left( y_{p_{i,t}^{X}} - y_{p_{j,t}^{Y}} \right)^2} \quad (1)$$

The above equation is also applied for the same object $(X = Y)$ and developed for two objects in two consecutive frames $t$ and $t' = t - 1$:

$$\phi^{X,Y}(i,j,t',t) = \left\| p_{i,t'}^{X} - p_{j,t}^{Y} \right\| \quad (2)$$

The joint distance features extracted from two objects are organized into a vector:

$$\Phi = \left[ \Phi^{X}(t) \quad \Phi^{Y}(t) \quad \Phi^{X,Y}(t) \quad \Phi^{X}(t',t) \quad \Phi^{Y}(t',t) \quad \Phi^{X,Y}(t',t) \quad \Phi^{Y,X}(t',t) \right] \quad (3)$$

where $\Phi^{X}(t) = \left\{ \phi^{X,X}(i,j,t) \right\}$, $\Phi^{Y}(t) = \left\{ \phi^{Y}(i,j,t) \right\}$, and $\Phi^{X,Y}(t) = \left\{ \phi^{X,Y}(i,j,t) \right\}$ for the same frame; and $\Phi^{X}(t',t) = \left\{ \phi^{X,X}(i,j,t',t) \right\}$, $\Phi^{Y}(t',t) = \left\{ \phi^{Y,Y}(i,j,t',t) \right\}$, $\Phi^{X,Y}(t',t) = \left\{ \phi^{X,Y}(i,j,t',t) \right\}$, and $\Phi^{Y,X}(t',t) = \left\{ \phi^{Y,X}(i,j,t',t) \right\}$ for consecutive frames.

**Joint direction:** The joint direction feature is identified as the angle between the vector $\vec{n}(i,j)$ and X-axis $\overrightarrow{Ox}$ as an illustration in Fig. 1. Generally, the direction feature between two joints $i$ of an object $X$ and $j$ of an object $Y$ within the frame $t$ is calculated as follows:

$$\theta^{X,Y}(i,j,t) = \angle\left( p_{i,t}^{X}, p_{j,t}^{Y} \right) \quad (4)$$

The above equation is also applied for the same object $(X = Y)$ and generally developed for two objects in two frame $t$ and $t' = t - 1$:

$$\theta^{X,Y}(i, j, t', t) = \angle\left(p^{X}_{i,t'}, p^{Y}_{j,t}\right) \quad (5)$$

The joint distance features extracted from two objects are organized into a vector:

$$\Theta = \left[\Theta^{X}(t) \quad \Theta^{Y}(t) \quad \Theta^{X,Y}(t) \quad \Theta^{X}(t', t) \quad \Theta^{Y}(t', t) \quad \Theta^{X,Y}(t', t) \quad \Theta^{Y,X}(t', t)\right] \quad (6)$$

where $\Theta^{X}(t) = \left\{\theta^{X,X}(i, j, t)\right\}$, $\Theta^{Y}(t) = \left\{\theta^{Y,Y}(i, j, t)\right\}$, and $\Theta^{X,Y}(t) = \left\{\theta^{X,Y}(i, j, t)\right\}$ for the same frame; while $\Theta^{X}(t', t) = \left\{\theta^{X,X}(i, j, t', t)\right\}$, $\Theta^{Y}(t', t) = \left\{\theta^{Y,Y}(i, j, t', t)\right\}$, $\Theta^{X,Y}(t', t) = \left\{\theta^{X,Y}(i, j, t', t)\right\}$, and $\Theta^{Y,X}(t', t) = \left\{\theta^{Y,X}(i, j, t', t)\right\}$ for the difference frames. Although, features are formed in distance and direction categories in (3) and (6), they can be structured in spatial and temporal relation as follows:

$$S = \left[\Phi^{X}(t) \quad \Phi^{Y}(t) \quad \Phi^{X,Y}(t) \quad \Theta^{X}(t) \quad \Theta^{Y}(t) \quad \Theta^{X,Y}(t)\right] \quad (7)$$

$$T = \left[\Phi^{X}(t', t) \quad \Phi^{Y}(t', t) \quad \Phi^{X,Y}(t', t) \quad \Phi^{Y,X}(t', t) \quad \Theta^{X}(t', t) \quad \Theta^{Y}(t', t) \quad \Theta^{X,Y}(t', t) \quad \Theta^{Y,X}(t', t)\right] \quad (8)$$

## 3 Experiment

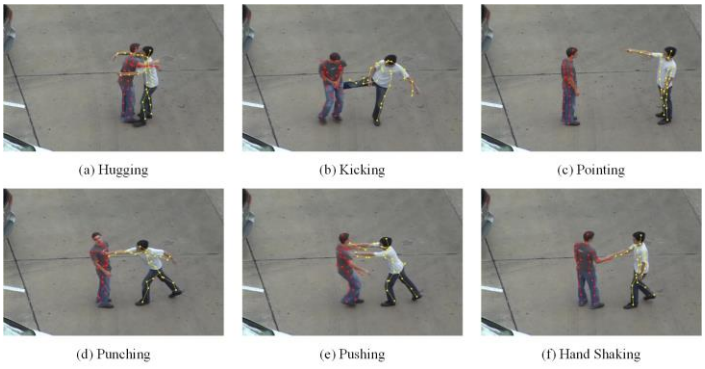


**Fig. 2.** UT-Interaction Dataset examples after using articulated joint estimation.

In the experiment stage, we benchmarked the method on the public UT-Interaction Dataset [7] that includes two sets of video data. Six interactive activities are presented in UT-dataset: hugging, kicking, pointing, punching, pushing, and hand shaking. Some snapshots of them with the articulated joint coordinates were represented in Fig.

2. For SVM classifier, the authors used LibSVM [9] with RBF kernel to solve the multiclass classification problem with 10-fold cross validation.

In the first experiment, we evaluated the proposed scheme on different feature categories. The classification results were represented through the confusion matrices in Fig. 3. Compared with the direction feature, distance feature reported a greater accuracy at most testing activities as in Fig. 3a and 3b. Furthermore, the temporal feature was also represent a more impressive results when compared with the spatial feature, shown in Fig. 3c and 3d., because of more useful interaction information contained in the temporal relation feature category. The highest accuracy was achieved when merging distance and direction feature in the consideration of spatio-temporal relation as Fig. 3e. In the second experiment, we make a comparison of performance in classification accuracy between the proposed method and existing approaches using the mean Accuracy (mAcc) and mean Average Precision (mAP). Based on the results in Table 1, the proposed method outperform others in the same experimental condition with 10-fold validation [3].

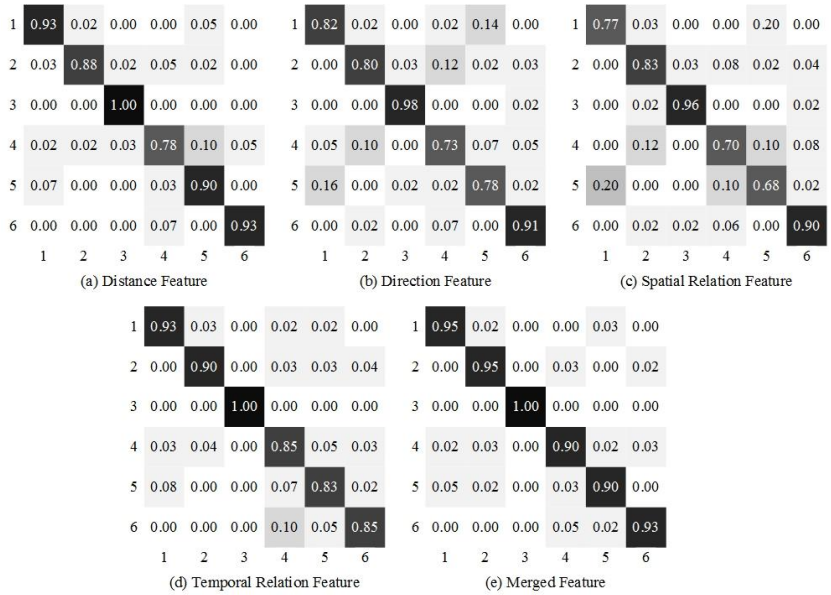


**Fig. 3.** Confusion matrices of different articulated-pose features. (1: Hugging, 2: Kicking, 3: Pointing, 4: Punching, 5: Pushing, 6: Hand shaking). The average classification rates are 90.56%, 83.89%, 80.83%, 89.44%, and 93.89% corresponding to the case of distance feature, direction feature, spatial relation feature, temporal relation feature and combination.

**Table 1.** Method comparison on UT-Interaction Dataset.

| Method | mAcc (%) | mAP (%) |
|---|---|---|
| PSR [2] | 49.09 | 45.90 |
| BoF [8] | 77.12 | 79.95 |
| SSR [3] | 87.42 | 91.81 |
| Our method | 93.89 | 93.89 |

## 4 Conclusion

In this work, we proposed an activity recognition method using the distance and direction features of all joint-pairs. From the joint coordinate dataset obtained by an articulate-pose estimation algorithm, features are extracted for two human objects in the spatio-temporal relation to fully describe the interactive activities. In the experiments, the authors evaluate the proposed method with different feature categories. In particular, our method is more efficient with the distance instead of the direction feature and with the temporal instead of the spatial feature. We extra compare our method with the state-of-the-art approaches to prove the impressive results is task of interaction recognition under the same validated condition.

**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) and also supported by the Industrial Core Technology Development Program, funded by the Korean Ministry of Trade, Industry and Energy (MOTIE), under grant number #10049079.

## References

1. Qiang Qiu, Zhuolin Jiang, Chellappa, R.: Sparse dictionary-based representation and recognition of action attributes. Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 707-714 (2011).
2. Pyry Matikainen, Martial Hebert, Rahul Sukthankar: Representing pairwise spatial and temporal relations for action recognition. European conference on Computer vision: Part I (ECCV'10), 508-521 (2010).
3. Lingxun Meng, Laiyun Qing, Peng Yang, Jun Miao, Xilin Chen, Metaxas, D.N.: Activity recognition based on semantic spatial relation. Pattern Recognition (ICPR), 2012 21st International Conference on, pp. 609-612 (2012).
4. Moghimi, M., Azagra, P., Montesano, L., Murillo, A.C., Belongie, S.: Experiments on an RGB-D Wearable Vision System for Egocentric Activity Recognition. Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on, pp. 611-617 (2014).
5. Yi Yang, Ramanan, D.: Articulated Human Detection with Flexible Mixtures of Parts. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.35, no.12, pp. 2878-2890 (2013).
6. Jia-xin Cai, Xin Tang, Guocan Feng: Learning Pose Dictionary for Human Action Recognition. Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 381-386 (2014).
7. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. Computer Vision, 2009 IEEE 12th International Conference on, vol., no., pp. 1593-1600 (2009).
8. Vincent Delaitre, Ivan Laptev, Josef Sivic: Recognizing human actions in still images: a study of bag-of-features and part-based representations, British Machine Vision Conference on, 97.1-97.11 (2010).
9. https://www.csie.ntu.edu.tw/~cjlin/libsvm/