HUMAN-HUMAN INTERACTION RECOGNITION BASED ON SPATIAL AND MOTION TREND FEATURE

Bangli Liu[1]*, Haibin Cai*[1]*, Xiaofei Ji*[2]*, Honghai Liu*[1]∗

[1]School of Computing, University of Portsmouth, UK
[2]School of Automation,Shenyang Aerospace University, China
Email: { bangli.liu, haibin.cai, honghai.liu }@port.ac.uk, jixiaofei7804@126.com

ABSTRACT

Human-human interaction recognition has attracted increasing attention in recent years due to its wide applications in computer vision fields. Currently there are few publicly available RGBD-based human-human interaction datasets collected. This paper introduces a new dataset for human-human interaction recognition. Furthermore, a novel feature descriptor based on spatial relationship and semantic motion trend similarity between body parts is proposed for human-human interaction recognition. The motion trend of each skeleton joint is firstly quantified into the specific semantic word and then a Kernel is built for measuring the similarity of either intra or inter body parts by histogram interaction. Finally, the proposed feature descriptor is evaluated on the SBU interaction dataset and the collected dataset. Experimental results demonstrate the outperformance of our method over the state-of-the-art methods.

***Index Terms*—** Human-human interaction, Action recognition, Semantic moving words, RGBD dataset.

1. INTRODUCTION

Human-human interaction recognition has been attracting increasing attention in computer vision field as its wide applications in visual surveillance, assistive living and human-machine interaction. However, it still remains a challenge due to mutual occlusion, various subject appearance or body size and complex context. Most existing human-human interaction recognition methods are based on RGB images [1–3]. These methods suffer from the clothing appearance variance, illumination variance, object scale changes and body parts localization difficulties.

The emergency of cost-efficient depth sensors gives us extra access to depth images and accurate body joints positions [4], which has motivated many RGBD-based works done for single person action recognition or human-object interaction recognition [4–8]. However, RGBD-based human-human interaction is not fully explored [9–13]. Yun *et al* [13] extracted features from sequences of skeleton joints and compared their recognition performance. These features include distance, joint movement, the geometric relationship between joints and planes, and velocity features. In [11], the relationship and motion information between interactive body pairs were used as distinctive features. Ji *et al* [9] associated the distance and motion features from single body part and interactive body part pairs for interaction representation. Apart from motion property, appearance features were also extracted from body parts in [10]. In these methods, the relation between interactive persons is mainly represented by the distance between body parts. The distance property is useful and discriminative, but only the distance might not be enough to reflect the inherent relations. Middle-level or high-level information is needed to reflect inherent characteristics of interactions.

In fact, several RGBD based single person action datasets such as *MSR Action3D* [4] and *MSRDailyActivity3D* [6] were also proposed for evaluation, while there are few RGBD-based human-human interaction datasets existing. To address these problems, a new large RGBD based human-human interaction dataset which consists of more interaction categories and samples is introduced. Furthermore, this paper proposes a feature descriptor called Spatial Relationship and Motion Trend Similarity(SRMTS), where semantic moving words are defined in 3D space and then the moving trend is quantified into the specific word. A Mercer kernel is built to measure the similarity between body parts by histogram intersection.

Our contributions can be summarized as follows: 1) The capture of aligned RGBD human-human interaction dataset. 2) The propose of viewpoint and location invariant feature descriptor Spatial Relationship and Motion Trend Similarity (SRMTS).

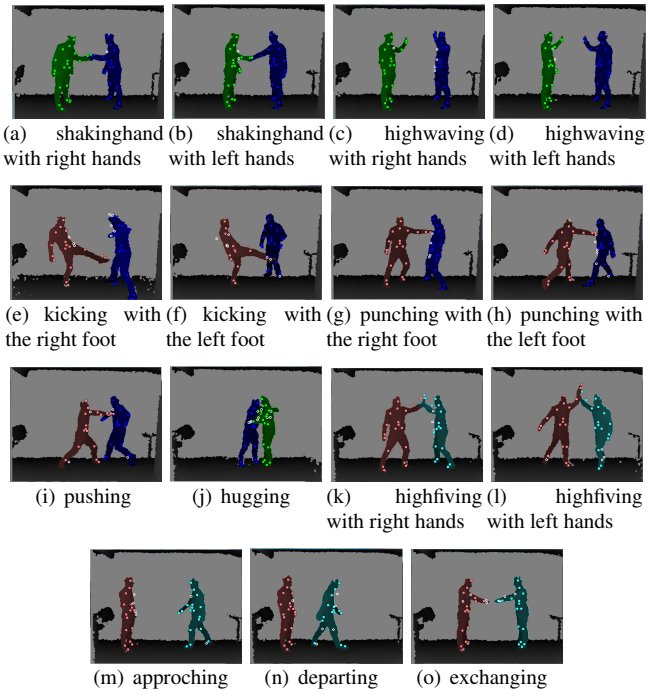
The rest of this paper is organized as follows: Section 2 introduces our dataset and compares it with the most existing datasets. Section 3 presents the proposed feature descriptor(SRMTS). Section 4 reports various experimental results as well as the comparison with the state-of-the-art methods. Section 5 summarizes the work of this paper.

2. KINECT BASED HUMAN-HUMAN INTERACTION DATASET

This section describes our human-human interaction dataset. This dataset contains 23 pairs of participants with various clothing color and body size. It has 10 human-human interaction categories: *shaking hands, high waving, kicking, punching, pushing, hugging, high-fiving, approaching, departing and exchanging objects*. Each category is repeated for three times and some of the categories might have more instances due to the distinguish between the right and left side. Thus, the total number of samples is around 900.

The dataset is collected using Kinect version 1 sensor. The recored data contains RGB data, original depth data, registered depth data and skeleton data. We further provide registered depth data to the RGB image which is useful for motion recognition when RGB and depth are jointly used in pixel level. The resolution of RGB and depth data is 640x480 and the dataset also provides 3D coordinates of 20 skeleton joints for each subject.

Compared to the existing datasets, our dataset has three advantages: 1) More interaction samples: our dataset has around 900 interactions, which is 3 times than that of [13]; 2) More complex: the performing habit of actors is considered by performing either the right or left side; 3) The registered depth image: the regristered depth information is useful for the segement of human body in RGB images and also provides convenience for jointly using RGB and depth data in pixel level.



(a) shakinghand with right hands (b) shakinghand with left hands (c) highwaving with right hands (d) highwaving with left hands

(e) kicking with the right foot (f) kicking with the left foot (g) punching with the right foot (h) punching with the left foot

(i) pushing (j) hugging (k) highfiving with right hands (l) highfiving with left hands

(m) approching (n) departing (o) exchanging

Fig. 1: Interaction samples of depth images and skeleton joints in Our dataset

3. PROPOSED METHOD

We extract spatial and motion trend relationship between body parts (e.g. left/right arm, left/right leg and head) to represent interactions. The configuration of body parts in the space is used to describe their spatial relationship. The moving direction of joints in each body part is firstly quantified into several semantic words and then a kernel is used to calculate the motion trend similarity by histogram intersection.

3.1. Spatial relationship

In this paper, each body part with four joints is described as follows: $P = \{b_1, b_2, b_3, b_4\}$, where b_i is $i-th$ joint, and it consists of three coordinates: $b_i = \{x_i, y_i, z_i\}$.

We apply the change of Euclidean distance between body part pairs to reflect the evolutional process of interactive human bodies. $Dist_{ij}$ is the distance between joint b_i and b_j:

$$Dist_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

Therefore, the spatial relationship between two body parts is represented as follows:

$$R\left(P_{p1}, P_{p2}\right) = \{Dist_{12}, ..., Dist_{ij}, ...\} \quad (2)$$

where $p1$ and $p2$ are body part numbers, i and j are joints from the same or different body part.

3.2. Motion trend similarity

3.2.1. Bags of semantic moving direction

The possible moving directions in 3D space are divided into 26 main semantic direction words, as shown in Fig.2: 6 axis directions(left, right, forward, backward, etc.) in black vectors,12 diagonal directions(leftforward, leftupward, rightward, rightbackward, etc.) in pink vectors and 8 gossip limit diagonal directions in blue vectors. We build the local coordinate system corresponding to the hip center: the $+x$ axis is the vector from the hip center to the right hip joint, and $+z$ is from the hip center to the spine joint, and the $+y$ axis is perpendicular to x and z and passes the hip center. As a consequence, the moving direction feature will be not effected by the change of viewpoint and location.

3.2.2. Semantic moving similarity between body parts

As the cosine value between two vectors can well measure the similarity of their direction, we use it to project the moving direction vector to one of 26 semantic direction words. The moving direction between two consecutive frames is defined as follows:

$$\mathbf{v_t^i} = \{x_t^i - x_{t-1}^i, y_t^i - y_{t-1}^i, z_t^i - z_{t-1}^i\} \quad (3)$$

where $\mathbf{v_t^i}$ is the moving direction, and x, y, z are three coordinates of joint i at time t or $t-1$. $\mathbf{v_t^i}$ is projected onto

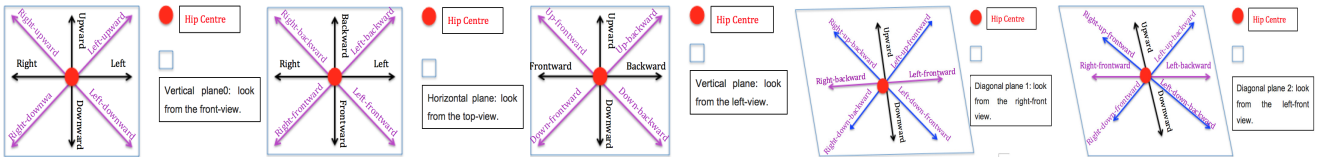
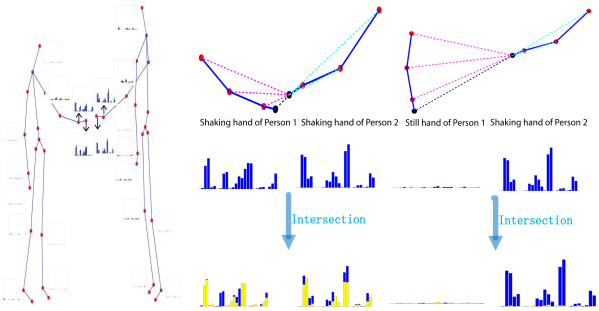


Fig. 2: Semantic moving directions. Black vectors are axis directions, pink vectors are diagonal directions, and blue vectors are Gossip limit diagonal directions.

the specific semantic word that is the closest to it. According to 26 semantic direction words, we can get a histogram Hi with 26 bins for joint i, as shown in Fig.3(a), and the value in each bin quantifies the motion degree in corresponding moving word. For the feature of one single body part, a concatenated moving direction vector with histograms of four ordered joints, is described as $P_n = \{H_i, i = 1, ..., 4\}$, n is the body part number.



(a) (b)

Fig. 3: (a) Quantization of moving directions in the space for each joint. (b) Semantic similarity between body parts by histogram intersection. The first row are the two active or inactive body parts during shaking hands. The blue histograms in the second row are the cumulative moving directions of joints (black ones) from each person, and the yellow part in the third row is the result of intersection, which reflect the degree of similarity between two body parts.

A kernel K is defined to calculate the similarity by intersecting the moving trend histograms. Histogram intersection [14] was firstly proposed for color indexing in object recognition and it can measure the degree of similarity between two histograms [15]. We use it to count the times of direction word in one joint that have corresponding times of the same direction word in another joint. Intra-similarity and inter-similarity are the relationship between body parts from the same person and from two persons, respectively.

We use the following intersection function to abtain the similarity between two corresponding histogram bins k for two joints (i and j):

$$InterSectionJ\left(b_i^k, b_j^k\right) = min\left(|b_i^m|, |b_j^m|\right) \tag{4}$$

In order to make sure the corresponding bin in two histogram having the same dimension, all interactions have been interpolated into the same numbers of frames(N). We now represent each histogram H with an $N \times m$-dimensional vector:

$$H = \left(\overbrace{1, ..., 1}^{H_1}, \underbrace{0, ..., 0}_{N-H_1}, \overbrace{1, ..., 1}^{H_2}, \underbrace{0, ..., 0}_{N-H_2}, ..., \overbrace{1, ..., 1}^{H_m}, \underbrace{0, ..., 0}_{N-H_m}\right) \tag{5}$$

Therefore, the intersection between two histogram is equal to the inner product between these two vectors:

$$K\left(H\left(i\right), H\left(j\right)\right) = \sum_{k=1}^{m} InterSectionJ\left(b_i^k, b_j^k\right) = H\left(i\right) \cdot H\left(j\right) \tag{6}$$

According to [15], K here is a Mercer's kernel. Fig.3(b) is the flowchart of intersection between two joints.

The intersection process is executed between joint pairs. Since the sum of the Mercer's kernels is also a Mercer's kernel [16], the intra-similarity kernel K_{intra} and inter-similarity K_{inter} are also Mercer's kernels. Fallowing Eq.6:

$$\begin{aligned} K_{type} &= \sum_{p=1}^{8}\sum_{q=1}^{8}\sum_{k=1}^{26} InterSectionJ\left(b_p^k, b_q^k\right) \\ &= \sum_{p=1}^{8}\sum_{q=1}^{8} K\left(H\left(p\right), H\left(q\right)\right) \end{aligned} \tag{7}$$

Where p and q are the body parts from the same person when K_{type} is K_{intra}, while p and q are the body parts from two interactive persons when K_{type} is K_{inter}.

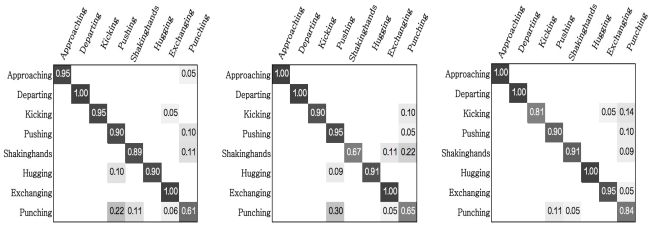
The final semantic moving direction similarity feature of body parts for each interaction$DirectionS$ is the concatenation of all body part pairs: $DirectionS = \{K_{intra1}, K_{inter1}, ...\}$, it consists of intra- and inter-similarity of all pairs of body parts.

3.3. SRMDS Feature Desriptor

In this paper, the spatial information and motion information are both extracted to represent different interaction categories. The effective feature descriptor named Spatial Relationship and Motion Trend Similarity(SRMTS) is the combination of spatial feature and semantic motion feature. A linear SVM [17] classification algorithm is used for interaction recognition.

4. EXPERIMENT

Inspired by [4], we collect different number of training samples to evaluate the performance of our feature descriptor: *Test One* (1/3 of the samples as training set, and the rest as testing set),*Test Two* (2/3 of the samples as training set and the rest as testing set) and *Cross Subjects Test* (half of the samples performed by half of interactive subject pairs as training dataset, and another half of the samples performed by another half of interactive subject pairs as testing data).



(a) *Test One* (b) *Test Two* (c) *Cross Subjects Test*

Fig. 4: Confusion Matixes on SBU interaction dataset

4.1. Experiment on SBU interaction dataset

Fig.4 shows the confusion matrixes of *Test One*, *Test Two*, and *Cross Subjects Test* on SBU interaction dataset. The confusion matrix reflects the confusion between different categories. It can be seen from Fig. 4 that *approaching* and *departing* can be easily recognized with 100% recognition rate due to their apparent motion, While the most common confusion is between *pushing* and *punching* in all three tests. The reason is that poses in *pushing* and *punching* are very similar.

We compare the recognition rate of our method to the results of the state-of-the-art: Joint features [13], CFDM [9] and [18] in Table 1. It indicates that the recognition rates of our method on *Test One*, *Test Two* and *Cross Subjects Test* are over 90% and the average rate is 91.12%, which improve the recognition rate by 10.82%, by 1.72% and by 4.22% than Joint Features, CFDM and [18] respectively. Fig. 5 gives the detailed recognition accuracy comparison of each category among Joint features [13], CFDM [9] and our method. Compared to Joint Features in [13], our method achieves better recognition on most of interactions, especially on *punching*, *hugging* and *exchanging*. Furthermore, the accuracies of most categories are higher than CFDM, apart from *shakinghands, hugging* and *exchanging*.

Table 1: Recognition Accuracy (%) on SBU dataset.

State-of-the-art	Joint features [13]	80.3
	CFDM [9]	89.4
	[18]	86.9
Proposed Method	*Test One*	90.58
	Test Two	90.28
	Cross Subjects Test	92.50
	Average	**91.12**

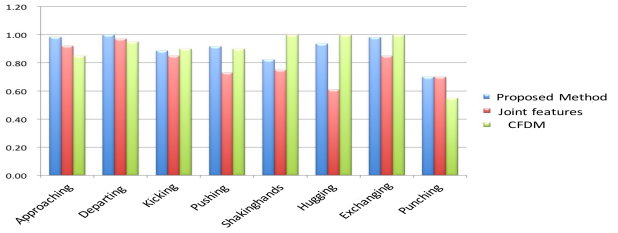
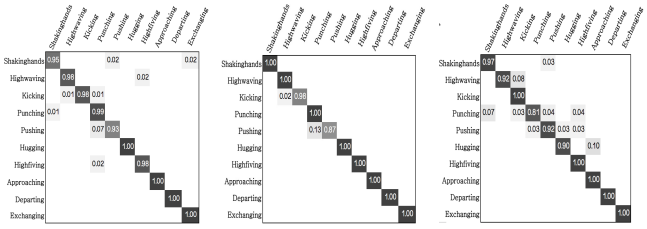


Fig. 5: Comparison of CFDM, Joint feature and Proposed Method by categories on SBU interaction dataset.

4.2. Experiment on our interaction dataset

For the newly collected dataset, we also use the *Test One*, *Test Two* and *Cross Subjects Test* for evaluation. Fig. 6 shows the confusion matrixes. With the proposed method, the recognition rates in most interaction categories are over 90%, and some reach to 100%. Although the similarity between interactions like *pushing* and *punching* is huge, the rates that *punching* is unexpectedly recognized as *pushing* and *pushing* is unexpectedly recognized as *punching* are very slow (0.03 and 0.04, respectively). Because the motion trend in the early stage of *hugging* is similar with that of *approaching*, the possibility of *hugging* recognized as *approaching* is relatively high (0.10)in *Cross Subjects Test*.



(a) *Test One* (b) *Test Two* (c) *Cross Subjects Test*

Fig. 6: Confusion Matixes on our dataset

5. CONCLUSION

In this paper, a new human-human interaction dataset with more samples is collected. Synchronized RGB images, depth images and skeleton joints with ground truth labels are provided. The feature descriptor(SRMTS) which combines spatial relationship and semantic motion trend similarity measured by a Mercer kernel is able to represent different interaction categories effectively. Experiment results on both SBU interaction dataset and our collected dataset demonstrate the outperformance of our method over the most of the state-of-the-art methods. As the skeleton joints will not be always available, our future work will focus on extracting more effective features by combining the information from depth images to achieve more accurate recognition.

Acknowledgements– This work is partially supported by the EU Seventh Framework Programme (No. 611391, Development of Robot-Enhanced therapy for children with AutisM spectrum disorders (DREAM)) and China Scholarship Council.

6. REFERENCES

[1] Fadime Sener and Nazli Ikizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 63–73, 2015.

[2] Snehasis Mukherjee, Sujoy Kumar Biswas, and Dipti Prasad Mukherjee, "Recognizing interactions between human performers by dominating pose doublet," *Machine Vision and Applications*, vol. 25, no. 4, pp. 1033–1052, 2014.

[3] Khai N Tran, Apurva Gala, Ioannis A Kakadiaris, and Shishir K Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognition Letters*, vol. 44, pp. 49–57, 2014.

[4] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 9–14, 2010.

[5] Omar Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 716–723, 2013.

[6] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1290–1297, 2012.

[7] Jiankun Hu, Weiye Zheng, Jih-Sheng Lai, Shuxi Gong, and Tao Xiang, "Exemplar-based recognition of human-object interactions," 2015.

[8] Bangli Liu, Hui Yu, Xiaolong Zhou, Dan Tang, and Honghai Liu, "Combining 3d joints moving trend and geometry property for human action recognition," in *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 000332–000337.

[9] Yanli Ji, Hong Cheng, Yali Zheng, and Haoxin Li, "Learning contrastive feature distribution model for interaction recognition," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 340–349, 2015.

[10] Rami Alazrai, Yaser Mowafi, and CS George Lee, "Anatomical-plane-based representation for human–human interactions analysis," *Pattern Recognition*, vol. 48, no. 8, pp. 2346–2363, 2015.

[11] Yanli Ji, Guo Ye, and Hong Cheng, "Interactive body part contrast mining for human interaction recognition," *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pp. 1–6, 2014.

[12] Tao Hu, Xinyan Zhu, Wei Guo, and Kehua Su, "Efficient interaction recognition through positive action representation," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[13] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 28–35, 2012.

[14] Michael J Swain and Dana H Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[15] Annalisa Barla, Emanuele Franceschi, Francesca Odone, and Alessandro Verri, "Image kernels," *Pattern Recognition with Support Vector Machines*, pp. 83–96, 2002.

[16] Francesca Odone, Annalisa Barla, and Alessandro Verri, "Building kernels from binary strings for image matching," *Image Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 169–180, 2005.

[17] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[18] Yanli Ji, Guo Ye, and Hong Cheng, "Interactive body part contrast mining for human interaction recognition," *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pp. 1–6, 2014.