Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 795360, 11 pages
http://dx.doi.org/10.1155/2013/795360

Research Article

Efficient Interaction Recognition through Positive Action Representation

Tao Hu,[1] Xinyan Zhu,[1,2] Wei Guo,[1] and Kehua Su[2]

[1] *Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China*
[2] *School of Computer Science, Wuhan University, Wuhan 430072, China*

Correspondence should be addressed to Wei Guo; guowei98032@gmail.com

Received 2 September 2013; Revised 11 November 2013; Accepted 13 November 2013

Academic Editor: Yue Wu

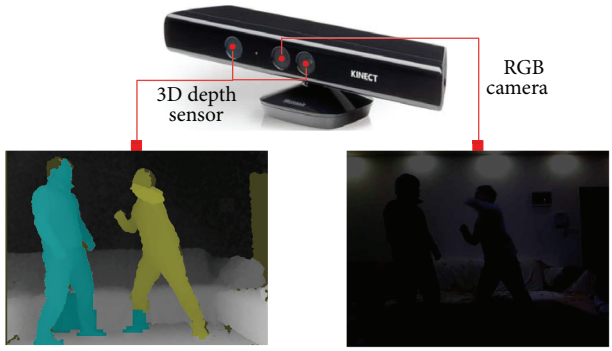
This paper proposes a novel approach to decompose two-person interaction into a Positive Action and a Negative Action for more efficient behavior recognition. A Positive Action plays the decisive role in a two-person exchange. Thus, interaction recognition can be simplified to Positive Action-based recognition, focusing on an action representation of just one person. Recently, a new depth sensor has become widely available, the Microsoft Kinect camera, which provides RGB-D data with 3D spatial information for quantitative analysis. However, there are few publicly accessible test datasets using this camera, to assess two-person interaction recognition approaches. Therefore, we created a new dataset with six types of complex human interactions (i.e., named K3HI), including kicking, pointing, punching, pushing, exchanging an object, and shaking hands. Three types of features were extracted for each Positive Action: joint, plane, and velocity features. We used continuous Hidden Markov Models (HMMs) to evaluate the Positive Action-based interaction recognition method and the traditional two-person interaction recognition approach with our test dataset. Experimental results showed that the proposed recognition technique is more accurate than the traditional method, shortens the sample training time, and therefore achieves comprehensive superiority.

1. Introduction

Over the last few decades, human activity analysis has undergone rapid development receiving increasing attention in many fields, such as intelligent surveillance, human-computer interaction, and elder care management [1, 2]. Human activity can be categorized according to complexity as partial body action [3], simple action [4], interaction activity [5, 6], or group activity [7]. Motivated by the activity classes drawn from [5, 6], this paper focuses on two-person interaction recognition of six complex interactions: kicking, pointing, pushing, punching, exchanging an object, and shaking hands.

Much research has been done on two-person interactions [5–10] with respect to the kinds of complex action relationships and human features necessary for recognition. For example, [5] took into account whether one person's hand is above another's shoulder or whether one person's foot is near another's torso. Reference [6] used head-pose, arm-pose, leg-pose, and overall body-pose estimation with both people for recognition. However, these processes are complex and time consuming and the recognition results might not be as accurate as required for a particular application. This paper proposes a new definition for interactions based on one person's behavior called Positive Action. In this method, one person's action plays the key role in an interaction; thus, two-person interaction recognition can be simplified into Positive Action recognition. This approach is simpler than traditional methods, saves computing time, and improves recognition results.

The recent proliferation of a cheap but effective depth sensor, the Microsoft Kinect [11], has created more opportunities for quantitative analysis of complex human activities. As compared to the traditional video camera, Kinect has the advantage of synchronous acquisition of color and depth images; with the use of depth maps, 3D information about a scene from a particular point of view is easily computed under diverse conditions [12]. This in turn will make behavior detection easier in badly lit or dark places. For example, Figure 1(a) represents a depth image captured by Kinect in



(a) Depth image (b) Color image (weak light)

FIGURE 1: RGB-D data captured by Kinect.

weak light, which clearly shows one person punching at another; Figure 1(b) shows a color image of this interaction synchronously captured with the depth image. With a traditional camera, only RGB images as seen in Figure 1(b) are collected, with limited value for surveillance and other applications. Unfortunately, there are few publicly accessible test datasets to assess two-person interaction recognition approaches using the depth sensor. Thus, we created a new dataset for two-person interaction. The first version of this original dataset is available to download on the Internet at http:// www.lmars.whu.edu.cn/prof_web/zhuxinyan/DataSetPublish/dataset.html.

The Microsoft Kinect sensor produces a new type of data, RGB-D data, which is an improvement on RGB images for human behavior recognition research. Therefore, many researchers have collected their own data and some of them are publicly accessible on the Internet [13–15]. In [16], Sung et al. produced a dataset including a total of twelve unique activities in five realistic domestic environments: office, kitchen, bedroom, bathroom, and living room. The RGBD-HuDaAct video database [17] collected in a lab environment includes 12 categories of human daily activities: making a phone call, mopping the floor, entering a room, and so forth. The LIRIS human activity dataset contains (gray/RGB/depth) videos showing people performing various activities taken from daily life (discussing, making telephone calls, exchanging an item, etc.); it includes information on not only the action class but also the spatial and temporal positions of objects in the video. However, these datasets only address individual activities and not two-person interactions [18].

Several more-than-one-person datasets were created using Kinect. In [19], the UT Kinect-human detection dataset was created: there are 98 frames with two people appearing in the scene at different depths in a variety of poses, including several simple interactions. In addition, [5] chose eight types of two-person interactions to establish another two-person dataset, including approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. However, this latter dataset is not publicly available on the Internet.

Depth imaging data produced by the Kinect sensor is driving new single and daily activity recognition problem research. For human activity or behavior representation, the method in [16, 20] detected and recognized different activities through body-pose features, hand position features, and motion information, using the Kinect sensor. In [17], Ni et al. proposed depth-extended feature representation methods to obtain superior recognition performance based on RGBD-HuDaAct datasets. Nowozin and Shatton [21] used skeletal features: joint velocities, joint angles, and joint angle velocities to reduce the latency in recognizing an action.

For human activity or behavior recognition, most efforts use HMM-based approaches. Park and Aggarwal [6] used HMMs for human motion recognition and combined it in a hierarchical way using DBNs (Dynamic Bayesian Networks). Vogler and Metaxas [22] presented parallel HMMs to recognize American sign language based on magnet tracking data, while Wilson and Bobick [23] proposed parametric HMMs to recognize human gestures. HMM-based recognition of more complex sequences is addressed by [24–26]. The method proposed in [24] was able to recognize motion units with optical flow data; in [25], Li proposed a landmark point trajectories-based approach to recognize view-invariant human actions and Chen et al. [26] presented a star skeleton model to recognize a single action and a series of actions.

Presently, there is little human interaction research based on Microsoft Kinect data and few papers report on a complex human activity dataset created to depict two-person interactions [5]. This research concluded that activity recognition represented by geometric relational features based on distance between all pairs of joints outperforms other feature choices. Our proposed approach and test dataset extend this research.

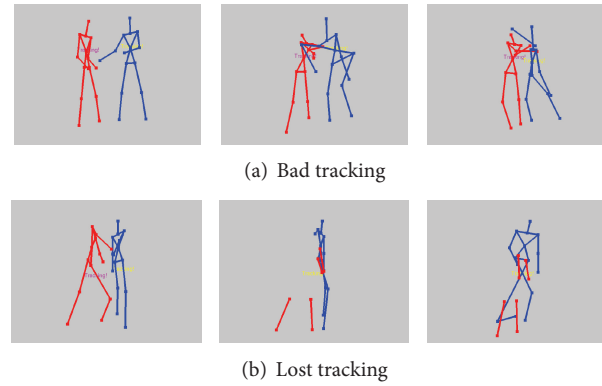
The contribution of this paper is twofold; we developed an efficient approach based on Positive Action representation to recognize two-person interactions and created a new dataset based on the Kinect sensor to test and verify methods. The rest of this paper is organized as follows. Section 2 shows our interaction dataset; Section 3 details the Positive Action definition and feature extraction method; Section 4 presents the Positive Action and the traditional interaction recognition method via HMMs; Section 5 demonstrates experimental results from two different approaches using our test dataset; finally Section 6 concludes this paper and discusses future work.

2. K3HI: Kinect-Based 3D Human Interaction Dataset

We collected two-person interactions using a Microsoft Kinect sensor. All videos were recorded in an indoor room while 15 volunteers performed activities. Each pair of people performed all types of interactions. The dataset has a total of approximately 320 interactions organized into eight categories. The first version of this dataset has been made publicly available to the research community to encourage progress in human action studies based on this new technology (http://www.lmars.whu.edu.cn/prof_web/zhuxinyan/DataSetPublish/dataset.html). Since approaching and departing activities are simple, recognition accuracy for both

interactions was almost 100% [5, 6]; therefore, we choose other types of relatively complex two-person interactions for recognition studies.

The most important data in our dataset is the spatial information (3D coordinates) of the two persons' skeletons. In order to ensure the integrity and continuity of target data, the original RGB images and depth information were ignored when capturing data. An articulated skeleton for each person was extracted using the OpenNI software [27] and Natural Interaction (NITE) Middleware provided by PrimeSense [28]. A skeleton was represented by the 3D positions of 15 joints, including head, neck, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, torso center, left hip, right hip, left knee, right knee, left foot, and right foot. However, when two persons overlapped, especially in a hugging activity (e.g., see Figure 2), full body tracking of interactions with NITE Middleware might be inaccurate. Bad and lost tracking will seriously affect interaction results, so hugging was not considered in our dataset. At last, six types of two-person interactions were captured, including kicking, punching, pointing, pushing, exchanging an object, and shaking hands. Figure 3 visualizes the collected interaction data as represented in the form of skeletons with different colors representing different actors.



(a) Bad tracking

(b) Lost tracking

Figure 2: Bad tracking and lost tracking for a hugging activity. (a) and (b) show the key process in hugging for two different pairs; the last two images in (a) represent bad tracking of human bodies, and (b) represents lost tracking of bodies.

3. Positive Action Representation

3.1. Positive Action Definition. Most existing work about human interactions focuses on two people, considering what kind of action relationship they have and what kind of features should be chosen to best represent an interaction [5, 6, 10–12]. Interactions can be classified into two groups: the first group indicates that one person acts first and the other person gives a responsive action, for example, kicking, pointing, punching, pushing, and so forth; the second group of interactions represents both people performing an almost identical synchronous action, for example, exchanging an object, shaking hands, and so forth. We propose that an interaction can be decomposed into a Positive Action and a Negative Action. For interactions in the first group, the person who acts first, resulting in the other person's reaction, performs a Positive Action. In the second group, since both people's behavior is similar and synchronized, we simply define the action, which moves with greater position changes in the first few frames, as the Positive Action. In all cases, a Negative Action is defined as a reciprocal action corresponding to a Positive Action in a two-person interaction.

After a Positive Action is identified, complex interaction recognition becomes relatively easy. Figures 4(a)–4(f) represent the original two-person interactions which were tested in [6], while Figures 4(a′)–4(f′) show the simplified results that the complex interactions are reduced into Positive Action-based representations. It can be seen that Positive Actions are discriminated with each other; therefore, only one person's features are taken into account and traditional interaction recognition can be transformed into Positive Action recognition.

3.2. Positive Action Extraction. Next, we obtained the Positive Actions in our dataset by means of mathematical analysis, especially for interactions in the first group as defined in Section 3.1. The window size for each interaction was approximately 25 frames. We only kept the first ten frames—since the action changes in the first few frames are enough to distinguish Positive Action and Negative Action. The extraction process for Positive Action is divided into the following three procedures.

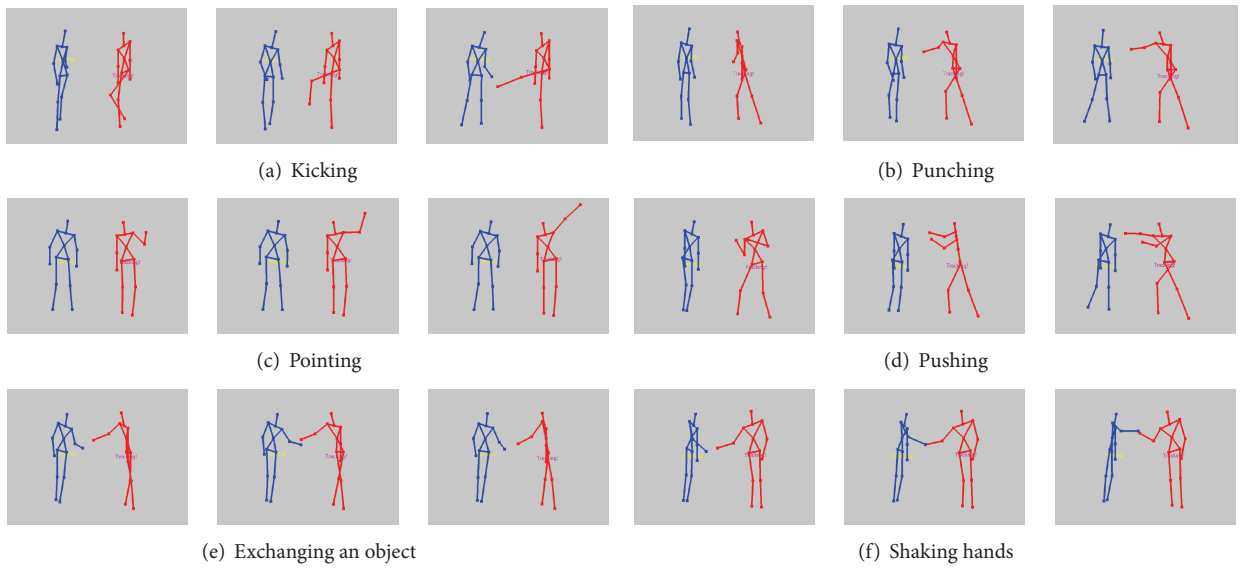
(1) Aligning the Sequence. For an interaction activity, there are always time or frame length variances when capturing the data. Before discerning a Positive Action, we first select the interactions of the same class to align the sequences. Then, the Dynamic Time Warping (DTW) model is used to align the sequences of the same activity class as mentioned in [29]. For each class, we selected a standard interaction sequence suitable for representation of the interaction process. We computed separately the minimal DTW distance between the remaining interaction sequences and the standard interaction sequence in the same class to find the optimal alignment.

In the DTW process, we express the feature vectors of two different sequences (in the same interaction class) as two time series (or frame series) $S_{T_1}^{(1)}$ and, $S_{T_2}^{(2)}$, defined as follows:

$$S_{T_1}^{(1)} = \left(s_1^{(1)}, s_2^{(1)}, \ldots, s_{t_1-1}^{(1)}, s_{t_1}^{(1)}\right),$$
$$S_{T_2}^{(2)} = \left(s_1^{(2)}, s_2^{(2)}, \ldots, s_{t_2-1}^{(2)}, s_{t_2}^{(2)}\right). \quad (1)$$

Accordingly, the costs between two series will be lower if they are similar, meaning that if two sequences are well aligned, the minimal DTW distance will be defined as

$$D\left(S_{T_1}^{(1)}, S_{T_2}^{(2)}\right) = \min\left\{D\left(S_{T_1-1}^{(1)}, S_{T_2-1}^{(2)}\right),\right.$$



(a) Kicking (b) Punching (c) Pointing (d) Pushing (e) Exchanging an object (f) Shaking hands

FIGURE 3: Skeleton visualization of interactions in our dataset. Three key poses were selected to represent the process of each interaction: (a) kicking, (b) punching, (c) pointing, (d) pushing, (e) exchanging an object, and (f) shaking hands.

$$D\left(S_{T_1-1}^{(1)}, S_{T_2}^{(2)}\right), D\left(S_{T_1}^{(1)}, S_{T_2-1}^{(2)}\right)\Big\} + d\left(S_{t_1}^{(1)}, S_{t_2}^{(2)}\right), \tag{2}$$

where $d(S_{t_1}^{(1)}, S_{t_2}^{(2)})$ is the feature distance at time t_1 and t_2 in two sequences $S_{T_1}^{(1)}$ and $S_{T_2}^{(2)}$.

It is known that there are two persons' 3D joint positions in an activity sequence, represented as

$$\left\{P_{(i,j)}^{(1)} = \left(x_{(i,j)}^{(1)}, y_{(i,j)}^{(1)}, z_{(i,j)}^{(1)}\right), P_{(i,j)}^{(2)} = \left(x_{(i,j)}^{(2)}, y_{(i,j)}^{(2)}, z_{(i,j)}^{(2)}\right)\right\}, \tag{3}$$

where $P_{(i,j)}^{(1)}$ and $P_{(i,j)}^{(2)}$ are the position set of the first and the second persons, respectively; i and j are the frame index and the joint index. We used the joint positions to characterize the feature in each frame for a distance computation between $S_{T_1}^{(1)}$ and $S_{T_2}^{(2)}$. The distance is described as

$$d\left(s_{t_1}^{(1)}, s_{t_2}^{(2)}\right) = \sum_{j=1}^{n}\left(\left|P_{(t_2,j)}^{(1)} - P_{(t_1,j)}^{(1)}\right|^2 + \left|P_{(t_2,j)}^{(2)} - P_{(t_1,j)}^{(2)}\right|^2\right), \tag{4}$$

where $|P_{(t_2,j)} - P_{(t_1,j)}|$ indicates the Euclidean distance at time t_1 and time t_2. Then, we placed the Euclidean distance into formula (4) to obtain the minimal DTW distance, finding the optimal alignment between variable length interaction sequences.

(2) Computing Key Joint Position Changes. We selected eight joints as key joints, which represent changes in the body's motion; these joints include the left and right elbow, left and right hand, left and right knee, and left and right foot.

The position changes of the joints were described by calculating the distances between neighboring frames, defined as follows:

$$D_{(i,i+1)}^{j} = \left|P_{i+1}^{(j;x,y,z)} - P_{i}^{(j;x,y,z)}\right|, \tag{5}$$

where $D_{(i,i+1)}^{j}$ is the Euclidean distance of a key joint j between frame i and $i+1$; $P_i^{(j;x,y,z)}$ indicates the position of joint j at frame i and (x, y, z) are the 3D coordinates.

(3) Identifying Positive Action. For actions in the first group which is defined in Section 3.1, it is tougher to extract Positive Action than it is in the second group. According to the benchmark in [30], human reaction time is around 0.2-0.3 s. Our collected data is 15 frames per second. When reaction time is converted into frames, it consists of 3-4 frames. This means that in the first group of interactions when a Positive Action starts, about 3-4 frames later, a corresponding Negative Action occurs.

In our Positive Action definition, because the joint positions in the first two adjacent frames change and conform to the benchmark, we can compare the maximum position changes of both persons' key joints between initial ith and $(i+3)$th frame of a sequence. The value of i for the standard interaction sequence mentioned in procedure (1) is one. For the other sequences after DTW processing, i will be different value. This is expressed as follows:

$$\text{Positive Action} = \arg\max\left(\max\left(D_{(i,i+3)}^{(p1;j)}\right), \max\left(D_{(i,i+3)}^{(p2;j)}\right)\right), \tag{6}$$

where $\max(D_{(i,i+3)}^{(p1;j)})$ and $\max(D_{(i,i+3)}^{(p2;j)})$ indicate the maximum position changes of joints for person one and person two

in an interaction; $\max(D_1, D_2)$ indicates that if $D_1 > D_2$, D_1 will represent the Positive Action and D_2 will represent the Negative Action; otherwise, D_2 will be the Positive Action. Figure 6 shows the processing results for Positive Actions, ignoring the Negative Actions. Each action has its own distinct characteristics, including easily confused interactions, such as exchanging an object and shaking hands.

Positive Action extraction is much easier in the second group as compared with the first group. According to the definition of Positive Action for group two, we also use (6); therefore, the person with the maximum $D^{(p;j)}_{(i,i+3)}$ performs the Positive Action.

In order to verify the method which is used to extract Positive Action, we selected the "kicking" action from the first group of interactions and "shaking hands" from the second group and calculated the position changes using (5) for the first 10 frames. Figure 5 shows the results: from Figure 5(a), it can be seen that as person one's right foot and right knee positions change from the first frame to the third frame, person two's left and right elbows as well as left and right hands positions also change in the fourth frame. These changes suggest that when person one starts to kick, person two's upper limbs react milliseconds later so that the first person's motion belongs to the Positive Action. However, Figure 5(b) does not show any connection between the two behaviors, except that both of their right hands and elbows move in a synchronized fashion. In general, experimental results support our Positive Action extraction method.

The visualization of Positive Actions is shown in Figure 6. Table 1 represents the extraction results for Positive Action with and without DTW for the first group, illustrating that the extraction results for Positive Action have greater accuracy after DTW preprocessing.

3.3. Feature Extraction. After Positive Actions are extracted, we utilize several body-pose features for motion-capture data representation and evaluate these features using our test dataset. One of the biggest challenges when using skeleton joints as a feature is that semantically similar motions may not necessarily be numerically similar [31]. To overcome this, [32] used relational body-pose features as introduced in [31], describing geometric relations between specific joints in a single pose or a short sequence of poses. Relational pose features were used to recognize daily-life activities performed by a single actor in a random forest framework; the features included joint, plane, and velocity features.

(i) Joint Features

Joint Distance. Let $p_{j,t} \in \mathfrak{R}^3$ be the 3D location of joint j in a Positive Action at time $t \in T$. The joint distance feature F_{JoiDis} is defined as the Euclidean distance between two joints at time t and is represented as

$$F^{(j_1,j_2;t)}_{\text{JoiDis}} = \left| p_{(j_1;t)} - p_{(j_2;t)} \right|, \tag{7}$$

where j_1 and j_2 are any two joints of a single person ($j_1 \neq j_2$).

Joint Motion. Similar to the joint distance feature, the joint motion feature F_{JoiMot} is defined as the Euclidean distance between joints j_1 at time t_1 and j_2 at time t_2. It captures Positive Action joint motions and is represented as

$$F^{(j_1,j_2;t_1,t_2)}_{\text{JoiDis}} = \left| p_{(j_1;t_1)} - p_{(j_2;t_2)} \right|. \tag{8}$$

(ii) Plane Features

Plane Feature. F_{Plane} captures the geometric relationship between a plane and a joint; F_{Plane} helps to express whether the left hand lies in front of the plane spanned by the right shoulder, left shoulder, or torso. It is defined as

$$F^{(j_1,j_2,j_3,j_4;t)}_{\text{Plane}} = \text{dist}\left(p_{(j_1;t)}, \left\langle p_{(j_2;t)}, p_{(j_3;t)}, p_{(j_4;t)} \right\rangle \right), \tag{9}$$

where $\langle p_{(j_2;t)}, p_{(j_3;t)}, p_{(j_4;t)} \rangle$ indicates the plane spanned by three other joints j_2, j_3, and j_4. $\text{dist}(p_{(j;t)}, \langle \cdot \rangle)$ represents the Euclidean distance from joint j to the plane.

Normal Plane Feature. F_{NorPlane} is similar to a plane feature; it helps to determine if and how far the joint "hand" is raised above the "shoulder"; F_{NorPlane} is defined as follows:

$$F^{(j_1,j_2,j_3,j_4;t)}_{\text{NorPlane}} = \text{dist}\left(p_{(j_1;t)}, \left\langle p_{(j_2;t)}, p_{(j_3;t)}, p_{(j_4;t)} \right\rangle \right), \tag{10}$$

where j_1 is the joint as in a plane feature and $\langle p_{(j_2;t)}, p_{(j_3;t)}, p_{(j_4;t)} \rangle$ indicates that the plane with normal vector $p_{(j_2;t)} - p_{(j_3;t)}$ passing through $p_{(j_4;k)}$. j_1, j_2, j_3, and j_4 represents different joints.

(iii) Velocity Features

Velocity Feature. F_{Vel} captures the velocity of one joint along a direction generated by two other joints at time t. F_{Vel} is defined as

$$F^{(j_1,j_2,j_3;t)}_{\text{Vel}} = \frac{v_{j_1;t} \cdot \left(p_{j_2;t} - p_{j_3;t} \right)}{\left| p_{j_2;t} - p_{j_3;t} \right|}, \tag{11}$$

where j_1, j_2, and j_3 are different joints.

Normal Velocity Feature. F_{NorVel} is similar to a normal plane feature; it captures the velocity of one joint along the direction of the normal vector of the plane generated by three other joints at time t. F_{NorVel} is defined as

$$F^{(j_1,j_2,j_3,j_4;t)}_{\text{NorVel}} = v_{j_1;t} \cdot \hat{n} \left\langle p_{j_2;t}, p_{j_3;t}, p_{j_4;t} \right\rangle, \tag{12}$$

where $\hat{n} \langle \cdot \rangle$ is the unit normal vector of the plane represented by $\langle \cdot \rangle$ when j_1, j_2, j_3, and j_4 are different joints.

Table 1: Accuracy of Positive Action extraction.

1st kind of interaction	Kicking	Pointing	Pushing	Punching
Accuracy (without DTW)	93.9%	95.8%	92.3%	90%
Accuracy (with DTW)	98.6%	99.2%	98.5%	97.7%

4. Positive Action Recognition via HMM

Hidden Markov Models (HMMs) are widely used for modeling time series data. Formally, a HMM can be described

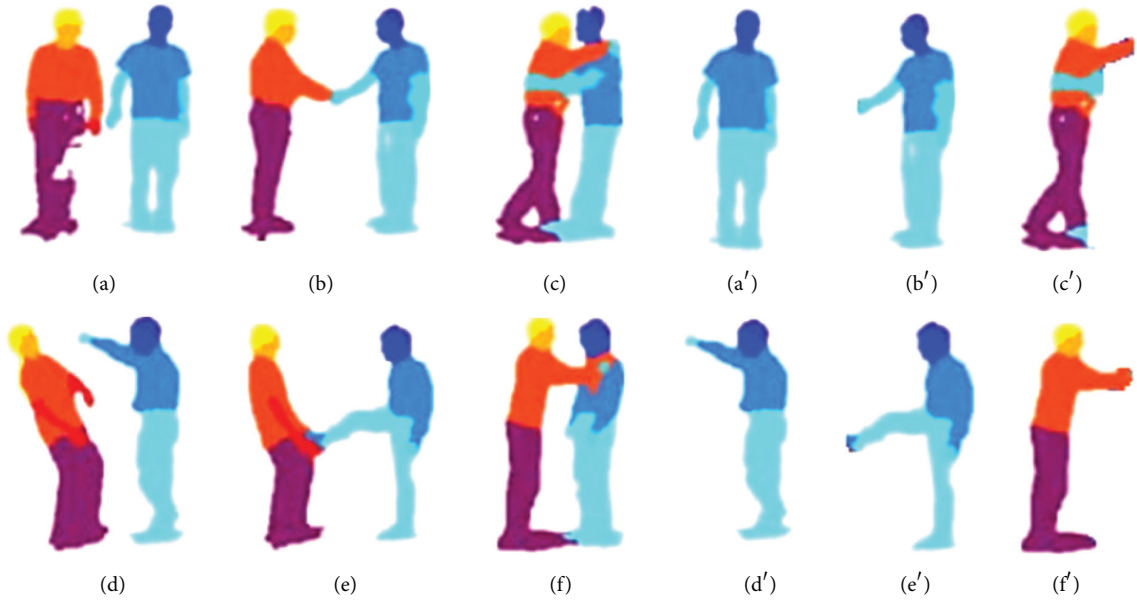


Figure 4: A comparison between interactions and Positive Actions. (a)–(f) show the original interaction data in [6] and (a′)–(f′) are the Positive Actions of one person described in this paper.

as a 5-tuple $\Omega = (\Phi, \Sigma, \pi, \delta, \lambda)$, where Φ are the hidden variables and δ are the transitions probabilities among states; these probabilities, as well as the starting probabilities π, are discrete. Every observation state has a set of possible emissions Σ and discrete/continuous probabilities λ for these emissions. A Gaussian Mixture Model (GMM) is used to represent the observation states for each hidden variable and to compute their probabilities [33]. GMM density is defined as the weighted sum of Gaussian densities.

In the training process, HMM parameters are initialized: we manually decided the observation states' number N and hidden states' number M; then we divided equally the data sequence into N parts and clustered each part using K-means to establish the GMM. After the HMM parameters are known, the Baum-Welch algorithm, also known as the Forward-Backward algorithm, was used to reevaluate the HMM parameters and to compute the output probability of observation sequence O_i^x (indicating the ith sample sequence of action x). Finally, the sequence probabilities are summed up and HMM parameters are confirmed until we get the maximum value $P(O \mid \Omega) = \sum P(O_i^x \mid \Omega_i)$. After training, we have six HMMs for each type of action.

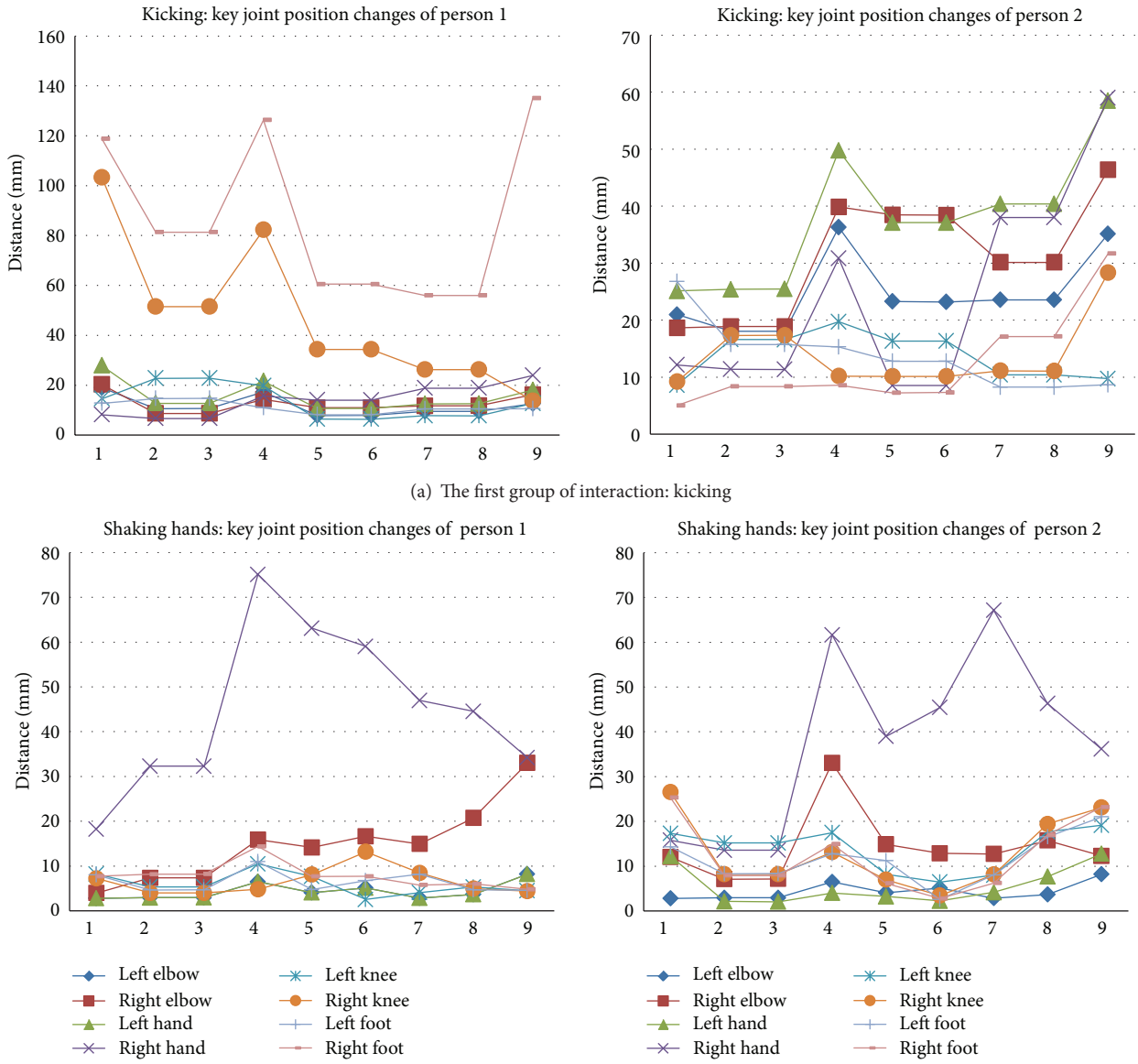
During the recognition process, given the data sequence of unknown action X, the feature vectors are extracted for each frame. Using the Viterbi algorithm, the likelihood $P_i = P(O_i^x \mid \Omega_i)$ of observation sequence O_i^X is generated. We repeated this procedure based on the six HMMs generated in training process and produced the probabilities p_i ($1 \leq i \leq 6$). Thus, by comparing the values p_i, we obtained the maximum likelihood $p_{\max}$, which represents the type of interaction.

5. Experimental Results

We selected the features extracted from among the Positive Actions identified in Section 3.2 to recognize interactions and used the features extracted from original interaction data as in [5]. Then, we compared and evaluated the recognition results from both approaches. The process for feature extraction and action recognition is illustrated in Figure 7.

In the Positive Action-based interaction approach, features as described in Section 3.3 were classified into three groups: joint features, plane features, and velocity features. In our experiments, we recognized six kinds of Positive Actions for each feature and mixed the features. There are fifteen joints (including 3D coordinates) for each action. Thus, the dimension of F_{JoiDis} is $C_{15}^2 = 105$ for each frame and the F_{JoiMot} was $C_5^2 \times C_T^2$ (T is the total number of frames for each interaction). Considering the larger dimensions of both plane and velocity features, we selected key joints to characterize the features. For plane features, the relationship between the four limbs and main body is critical; therefore, the plane was spanned from seven joints ("head," "neck," "left shoulder," "right shoulder," "torso," "left hip," and "right hip") and eight joints for the target joint. In this way, we created a lower dimension $C_7^3 \times 8$ for each frame. However, the feature dimensions were larger than the training sample number; thus, Principal Component Analysis (PCA) was used to reduce the dimensions.

To classify interactions, evaluation is done with a 4 fold cross-validation: 3 folds are used for training and 1 for testing. Based on the fact that the 3 state HMM performs much better than the 4- and 5 state HMMs in our experiments, we trained a 3 state, continuous HMM with GMM. As expected,



(a) The first group of interaction: kicking

(b) The second group of interaction: shaking hands

Figure 5: Key joints position changes in two groups of interactions during the first 10 frames. (a) shows the first group of interaction with "kicking" as an example; (b) shows the second group and takes "shaking hands" as an example.

the transition probabilities and the observation probabilities turned out to be different for different actions. After training, the HMM parameters are known while the Viterbi algorithm was used to find the maximum likelihood category. Table 2 shows the experimental results for each kind of feature representation.

For the traditional two-person relationship-based interaction recognition method (called the old approach in the rest of this paper), three kinds of features referring to [5] were also extracted based on the original captured data (see Figure 3). The training and recognition process was identical with the Positive Action-based (new) method. Figure 8 shows the recognition results in a confusion matrix: (a)–(c) represents the Positive Action-based approach and (d)–(f) for the values

Table 2: Interactions recognition results via Positive Action-based representation.

Features	Average accuracy
Raw position	45.2%
Joint distance	76.1%
Joint motion	75.6%
Plane	63.2%
Normal plane	65.3%
Velocity	44.2%
Normal velocity	41.2%

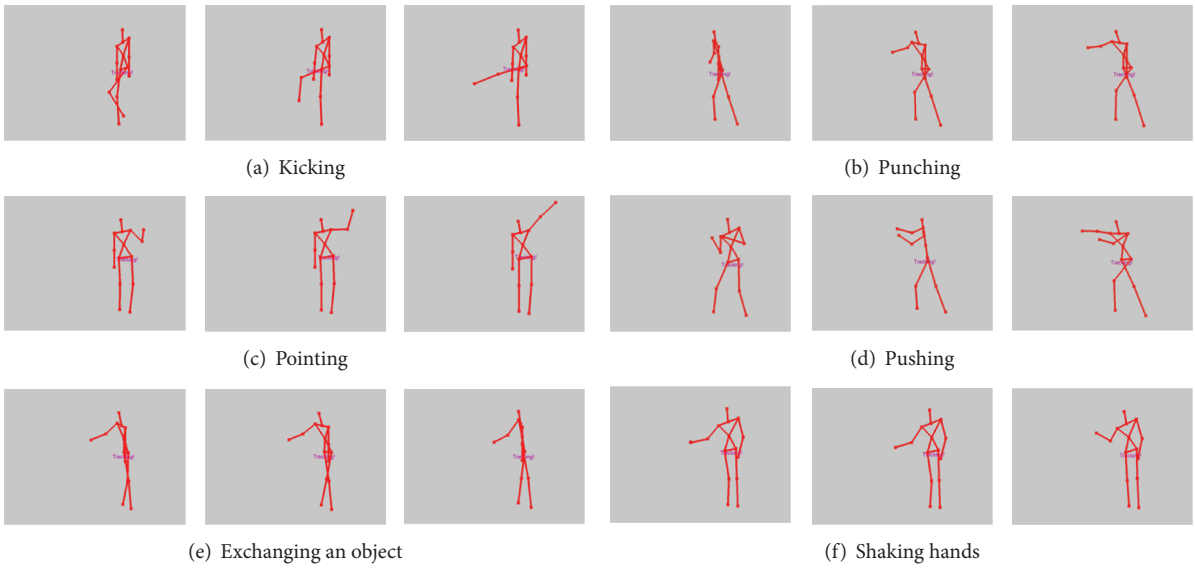


Figure 6: Skeletons visualization of Positive Actions. The red skeletons show only Positive Actions in two-person interactions. These are considered as the interaction representation and Negative Actions are ignored in the recognition process.

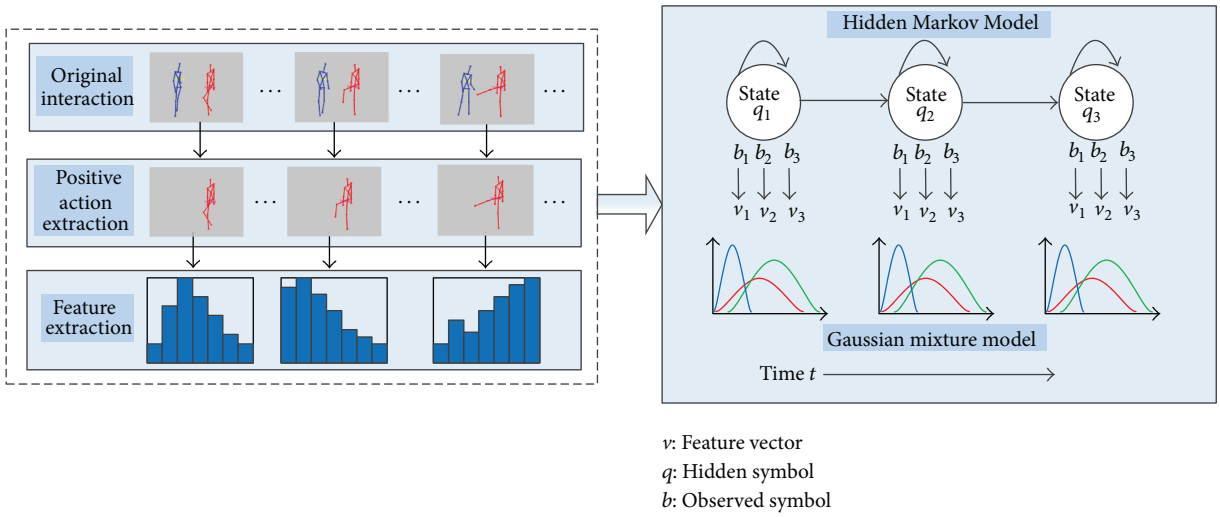


Figure 7: Flow of the interaction recognition system.

generated by the old approach. The confusion matrix also compares different kinds of features for recognition: joint features include the joint motion and joint distance features; plane features include the plane and normal plane features; velocity features include the velocity and normal velocity features. The average recognition accuracy for each kind of feature from (a) to (c) is 78.67%, 66.83%, and 55.67%; the average accuracy from (d) to (f) is 70.00%, 61.67%, and 48.67%. Therefore, joint features-based recognition results are better than plane and velocity features, suggesting that geometric relational features based on the distance between joints outperform other feature choices, verifying the conclusions found in [5]. Furthermore, in both the old and new approaches, there exists some confusion between "pointing" and "punching" and between "exchanging an object" and "shaking hands." Our results show that these actions are similar, leading to lower recognition accuracy.

Most importantly, the average accuracy for interaction recognition based on Positive Action representation, as proposed in this paper, is 7% greater than two-person relationship-based approaches, especially since geometric relational feature-based recognition is almost 10% greater. There are several reasons for these results. First, a two-person feature representation is more complex than a Positive Action-based representation, creating unstable factors. For example, the "pointing" interaction in normal plane features: the Positive Action-based method only judges whether one person's "hand" position is higher than its' own "shoulder";

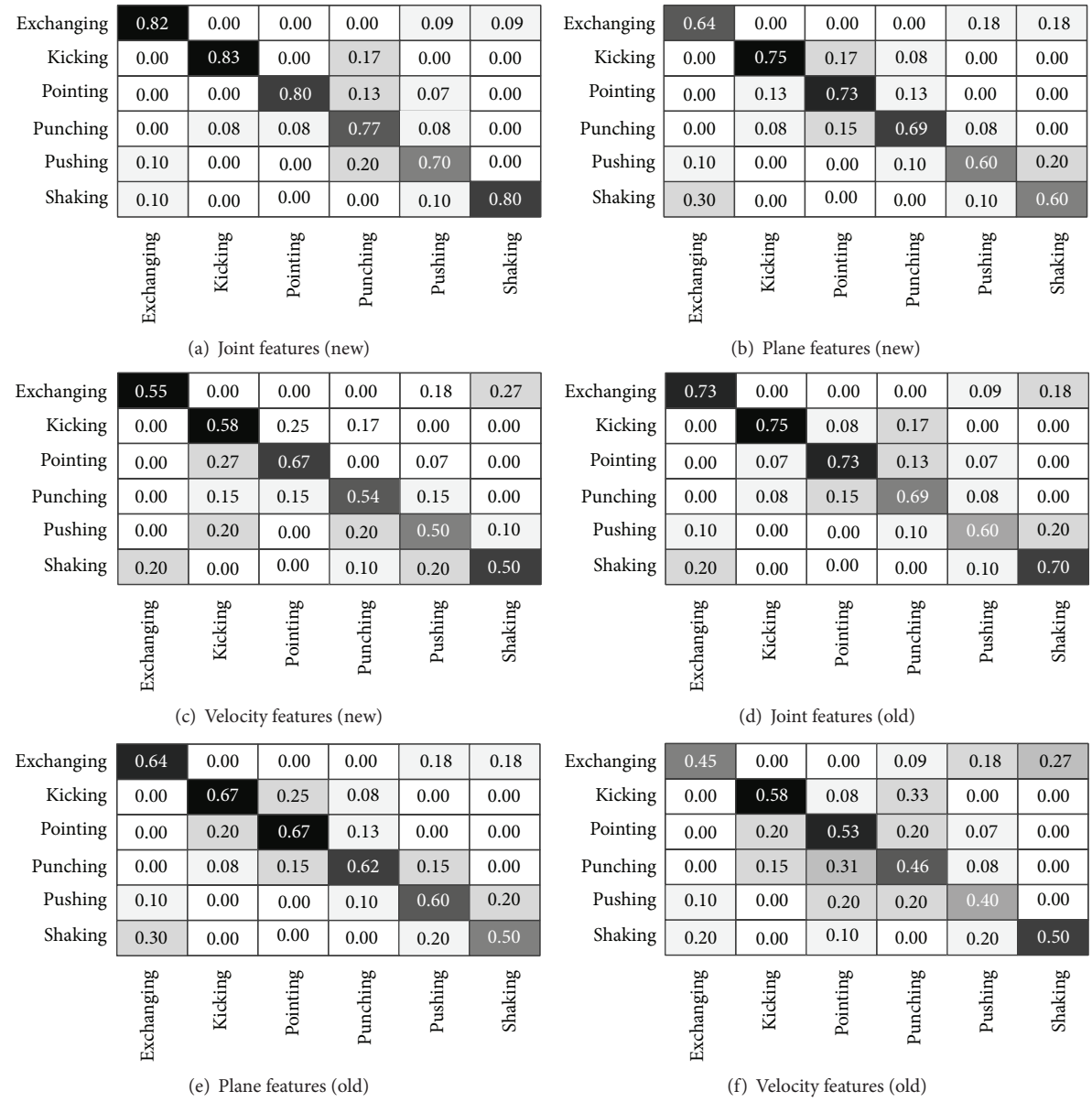


Figure 8: Confusion matrix for different features for two approaches. (a)–(c) are the recognition results based on Positive Action representation; (d)–(f) are the results based on old one referred to [5].

however, the old approach as in [5] must judge the spatial relationship for both persons' shoulders, which will lead to more conditions for recognition; therefore, the Positive Action-based approach needs less training samples than the old approach to get more or less the same recognition accuracy. Second, for the same kind of feature, the Positive Action-based representation method has fewer dimensions than the old approach. The old approach therefore is more sensitive during dimension reduction in the training process; thus, its recognition accuracy will be lower.

To verify the generalizability of our proposed method, we tested the dataset against two more classifiers, including Support Vector Machines (SVMs) and Multiple Instance Learning (MIL). The test features were represented by the combination of joint distance and joint motion. The results in Table 3 suggest that MIL has better performance than SVMs while the Positive Action-based method is much better than the two-person based method. Therefore, different classifier supports the conclusion that our new method is effective.

In addition to a comparison of the interaction recognition accuracy for both approaches, we also compared time costs and evaluated the training time to arrive at optimal HMM parameters (see Figure 9). The average training time for three kinds of features based on Positive Action representation is 42.47 MS (millisecond), 79.52 MS, and 67.88 MS, while for the old approach referring to [5], the average training time is 63.27 MS, 199.6958 MS and 156.3827 MS.

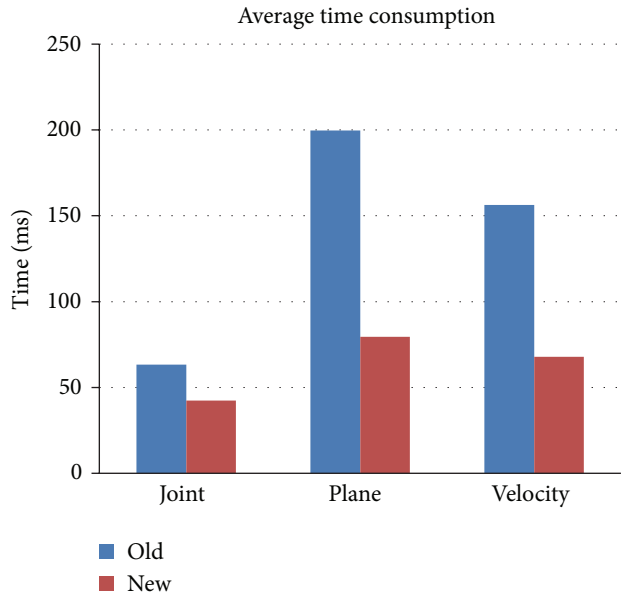


Figure 9: Average time cost for training samples. It is the old and new methods that are evaluated according to three kinds of features: joint, plane, and velocity features.

The Positive Action-based representation method consumes less time than the old approach.

In summary, Positive Action-based representation for two-person interaction recognition outperforms the old approach; not only is its recognition accuracy better, but also the time cost for training is less. So, the new method transforms a relatively complex two-person interaction into a simpler Positive Action, making the recognition procedure more cost effective while maintaining or even improving recognition quality. Therefore, the new proposed approach is efficient for interaction recognition.

6. Conclusion

This paper presented a novel approach to recognize relatively complex human interactions: different from many existing interaction recognition methods, we focused our research on single actions which are useful when distinguishing differences between types of interactions. Two-person interaction recognition is transformed into Positive Action-based recognition.

The key contributions of this paper are as follows: (1) we investigated the reciprocal relationships in two-person interaction and proposed a new definition for single person's behavior called Positive Action; (2) two-person interactions were recognized based on Positive Action representation via continuous HMMs; (3) a new test interaction dataset based on Microsoft Kinect camera was created and it is publicly available; our experimental results demonstrate that the proposed method outperforms old approaches based on two-person relationships.

In the future, we plan to find more volunteers to capture more data and extend our interaction dataset to include additional interaction categories. More importantly, owing to the limitations of human tracking software, such as the NITE Middleware or the Windows SDK for Kinect, there occasionally are some inaccurate tracking results. Therefore, we need to find a better way to track human actions, further improving the recognition accuracy.

Table 3: The performance on more classifiers.

Classifier	Positive Action	Two persons
SVMs	81.67%	76.67%
MIL	83.33%	78.33%

Acknowledgments

The authors are grateful to the volunteers for capturing data. This work was supported by the National Natural Science Foundation (41301517), the National 863 Key Program (2013AA122301), the National Key Technology R&D Program (2012BAH35B03), Chinese NSF Creative Research Group project (41023001), the National 973 Program (2011CB707001), the Fundamental Research Funds for the Central Universities (2012619020215), and Doctoral Fund of Ministry of Education (20120141120006).

References

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys*, vol. 43, no. 3, article 16, 2011.

[2] R. W. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] J. Yang, Y. F. Li, K. Wang, Y. Wu, G. Altieri, and M. Scalia, "Mixed signature: an invariant descriptor for 3D motion trajectory perception and recognition," *Mathematical Problems in Engineering*, vol. 2012, Article ID 613939, 29 pages, 2012.

[4] S. Sadek, A. A. Hamadi, B. Michaelis, and U. Sayed, "Chord-length shape features for human activity recognition," *ISRN Machine Vision*, vol. 2012, Article ID 872131, 9 pages, 2012.

[5] Y. Kiwon, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proceedings of the IEEE Computer Society Computer Vision and Pattern Recognition Workshops (CVPRW '12)*, vol. 28, no. 35, pp. 28–35, Providence, RI, USA, June 2012.

[6] S. Park and J. K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Systems*, vol. 10, no. 2, pp. 164–179, 2004.

[7] L. S. Wu, L. M. Xia, and D. Y. Luo, "Survey on human interactive behavior recognition and comprehension," *Computer Applications and Software*, vol. 28, no. 11, pp. 60–63, 2011.

[8] K. Sato and J. K. Aggarwal, "Tracking and recognizing two-person interactions in outdoor image sequences," in *Proceedings of the IEEE Workshop on Multi-Object Tracking*, pp. 87–94, Vancouver, Canada, 2001.

[9] S. Park and J. K. Aggarwal, "Recognition of human interaction using multiple features in gray scale images," in *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 1, pp. 51–54, Barcelona, Spain, 2000.

[10] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 1709–1718, New York, NY, USA, June 2006.

[11] http://www.microsoft.com/en-us/kinectforwindows/.

[12] S. Escalera, "Human behavior analysis from depth maps," in *Articulated Motion and Deformable Objects*, vol. 7378 of *Lecture Notes in Computer Science*, pp. 282–292, Springer, Berlin, Germany, 2012.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2556–2563, Barcelona, Spain, November 2011.

[14] M. S. Ryoo and J. K. Aggarwal, "ICPR contest on Semantic Description of Human Activities (SDHA)," UT-Interaction Dataset, 2010.

[15] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, December 2012.

[16] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proceedings of the AAAI Workshop on Pattern, Activity and Intent Recognition (PAIR '11)*, pp. 47–55, San Francisco, Calif, USA, August 2011.

[17] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: a color-depth video database for human daily activity recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW '11)*, pp. 1147–1153, Barcelona, Spain, November 2011.

[18] C. Wolf, J. Mille, E. Lombardi et al., "The LIRIS human activities dataset and the ICPR, 2012 human activities recognition and localization competition," Tech. Rep. RR-LIRIS-2012-004, LIRIS Laboratory, Lyon, France, 2012.

[19] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '11)*, pp. 15–22, Colorado Springs, Colo, USA, June 2011.

[20] S. Z. Masood, C. Ellis, A. Nagaraja, M. F. Tappen, J. J. Laviola, and R. Sukthankar, "Measuring and reducing observational latency when recognizing actions," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW '11)*, pp. 422–429, Barcelona, Spain, November 2011.

[21] S. Nowozin and J. Shotton, "Action points: a representation for low-latency online human action recognition," Tech. Rep., Microsoft Research, 2012.

[22] C. Vogler and D. Metaxas, "Parallel hidden Markov models for american sign language recognition," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV'99)*, vol. 1, pp. 116–122, Kerkyra, Greece, September 1999.

[23] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, 1999.

[24] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz, "HMM-based human motion recognition with optical flow data," in *Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots*, pp. 425–430, Paris, France, December 2009.

[25] X. Li and K. Fukui, "View invariant human action recognition based on factorization and HMMs," *IEICE Transactions on Information and Systems*, vol. 91, no. 7, pp. 1848–1854, 2008.

[26] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks (VSSN '06)*, pp. 171–178, New York, NY, USA, October 2006.

[27] http://www.openni.org/.

[28] http://www.primesense.com/.

[29] S. Sempena, N. U. Maulidevi, and P. R. Aryan, "Human action recognition using dynamic time warping," in *Proceedings of the International Conference on Electrical Engineering and Informatics (ICEEI '11)*, pp. 1–5, Bandung, Indonesia, July 2011.

[30] http://www.humanbenchmark.com.

[31] M. Muller, T. Roder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 677–685, 2005.

[32] A. Yao, J. Gall, G. Fanelli, and L. V. Gool, "Does human action recognition benefit from pose estimation?" in *Proceedings of the British Machine Vision Conference (BMVC '11)*, pp. 67.1–67.11, University of Dundee, Dundee, UK, September 2011.

[33] B. Poonam, K. Anuj, K. Sumit, S. Akash, and G. Shitij, "Improved hybrid model of HMM/GMM for speech recognition," in *Book 5 Intelligent Technologies and Applications*, Institute of Information Theories and Applications, Sofia, Bulgaria, 2008.