applied sciences

MDPI

Article

A New Framework of Human Interaction Recognition Based on Multiple Stage Probability Fusion

Xiaofei Ji [1],*, Changhui Wang [1] and Zhaojie Ju [2]

[1] School of Automation, Shenyang Aerospace University, Shenyang 110036, China; wangchanghui80@126.com
[2] School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK; zhaojie.ju@port.ac.uk
* Correspondence: jixiaofei7804@126.com; Tel.: +86-024-8972-4448

Academic Editors: Plamen Angelov and José Antonio Iglesias Martínez
Received: 19 February 2017; Accepted: 24 May 2017; Published: 1 June 2017

Abstract: Visual-based human interactive behavior recognition is a challenging research topic in computer vision. There exist some important problems in the current interaction recognition algorithms, such as very complex feature representation and inaccurate feature extraction induced by wrong human body segmentation. In order to solve these problems, a novel human interaction recognition method based on multiple stage probability fusion is proposed in this paper. According to the human body's contact in interaction as a cut-off point, the process of the interaction can be divided into three stages: start stage, execution stage and end stage. Two persons' motions are respectively extracted and recognizes in the start stage and the finish stage when there is no contact between those persons. The two persons' motion is extracted as a whole and recognized in the execution stage. In the recognition process, the final recognition results are obtained by the weighted fusing these probabilities in different stages. The proposed method not only simplifies the extraction and representation of features, but also avoids the wrong feature extraction caused by occlusion. Experiment results on the UT-interaction dataset demonstrated that the proposed method results in a better performance than other recent interaction recognition methods.

Keywords: human interaction recognition; piecewise fusion; weighted fusing; Hidden Markov Model

1. Introduction

Human interaction recognition and understanding is an important research topic in the computer vision community [1,2]. This research direction has received considerable attention in recent years due to the increasing number of potential applications, such as visual surveillance, human–computer interaction, video indexing and retrieval, smart homes/offices, healthcare rooms, physical sciences, health-related issues, natural sciences and industrial academic areas [3–6], etc.

The human interaction recognition requires an understanding of spatio–temporal relationships between different objects, in additional to individual variability, cluttered background, viewpoint changes, and other environment induced conditions [7]. So it remains challenging to recognize human interaction from complex scenes. Recent research on interaction recognition can be characterized by two classes of methods.

(1) **Interaction is recognized as a general action.** This kind of method usually represents the interaction as an integral descriptor including all the people involved in the interaction. Then a traditional classier is utilized to classify interactions. Yu et al. obtained a powerful discriminative codebook by introducing semantic texton forests (STFs) to local space–time volumes. Then the hierarchical k-means algorithm with pyramid match kernel is applied to achieve robust structural matching and interaction recognition [8]. The feature extraction of this method is relatively simple, however, the matching method is relatively complicated. Burghouts et al. improved the spatio–temporal representation by introducing a spatio–temporal layout of actions and obtained

successful human interaction recognition [9]. Peng et al. utilized multi-scale dense trajectories with a four advanced feature (DT (Dense Trajectory) shape, Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF), Motion Boundary Histograms(MBH)) encoding method to achieve human interaction recognition [10]. Li et al. proposed a hybrid framework which incorporates both a global feature (Motion Context) and a local feature (spatio–temporal interest point (STIP)) to recognize human interactions. The method achieves promising results by respectively using a Genetic Algorithm (GA)-based random forest and calculating the S–T correlation score as the recognition method [11]. As Kinect was introduced by Microsoft, both RGB images and depth images of the scene can be simultaneously captured. Some researchers have studied the RGBD-based human interaction recognition method. Ni et al. combined color and depth information to develop two feature representations, i.e., spatio–temporal interest points (STIPs) and motion history images (MHIs). The proposed multi-modality fusion method was tested and demonstrated superior performance on a home-monitoring oriented human interaction recognition dataset [12]. Yun et al. designed a variety of related distance features (such as joint distance, joint movement, plane features, normal plane features, velocity features and normal velocity features) to carry out a two-person interaction recognition [13]. This method can achieve real-time human interaction detection, but the performance of the method depends on the accurate extraction of the joint point.

This kind of method treats people as an entity and so does not need to segment the individual as a feature in the interaction, which makes the processing method relatively simple. However, this kind of method does not accurately represent the intrinsic properties of the interaction, and therefore, a better performance requires comprehensive motion features and a matching method.

(2) **Interaction recognition using motion co-occurrence.** This kind of method proposes that the interaction is composed of a set of temporal-ordered elementary actions performed by the different persons involved in the interaction [14,15]. Kong et al. proposed interactive phrases to describe the motion relationships between two people in the interaction. Then, a discriminative model is proposed to encode interactive phrases based on the latent Support Vector Machine (SVM) formulation [16]. The interactive phrases consider more detail in the interaction, so the recognition accuracy is greatly improved, but it does require that all possible rules are predefined. SLIMANI et al. proposed a co-occurrence of a visual words method for human interaction recognition [17]. The method represents the interaction between persons by calculating the number of times that visual words occur simultaneously for each person involved in the interaction. While the implementation of this method is simple, the co-occurrence relationships are not expressive enough to effectively deal with interactions that contain large variations [18,19]. In general, this kind of method can achieve more accurate and robust results by exploiting rich contexture information in human interactions. However, the recognition results always depend on the accurate feature segmentation of the individual and the stability of the individual behavior model.

Based on the above analysis, most of the current interactive behavior recognition methods always address the interaction as a whole to extract features, or separate the interactive behavior into two independent objects to extract features. Obviously, these two kinds of methods cannot reasonably represent the dynamic development process of interactive behavior. A new framework of interactive behavior recognition is therefore proposed by combining those two methods in this paper, as shown in Figure 1. The proposed framework includes the following modules:

(1) **Video segmentation in time domain.** Interactive behavior is divided into three stages, i.e., start stage, middle stage and end stage in accordance with the distance between two persons.

(2) **Region of interest extraction.** The interaction areas are divided into two independent regions at the start and end stages, when the distance between two persons is extensive. The interaction area is extracted as a whole in the middle stage, when the distance between two persons is smaller and the participants are close to each other or in contact.

(3) **Feature extraction.** The HOG (Histogram of Oriented Gradient) descriptor is used to represent the region of interest (ROI) of each frame. The HOG descriptors are separately extracted in two

ROIs at the start and end stages, and the HOG descriptor is extracted in the global ROI in the middle stage.

(4) **Interaction model training.** By using HOG features which have been extracted from each interaction stage. The Hidden Markov model (HMM) is chosen as our action model, because it captures the dynamics of the human action and has robustness in relation to noise.

(5) **Interaction recognition.** When the unknown interaction is taking place, the video segmentation and feature extraction are first performed. Then the probabilities of the test sequence for the sub-HMMs are computed. At the end, the recognition result is obtained by weighted fusing of the likelihood probability of these three stages.

Interaction is recognized as a general action method that can fully express global information, and interaction recognition using motion co-occurrence can accurately describe individual behavior, so the proposed method of combining those two kinds of methods is more effective for representing the dynamic process of interactive behavior.

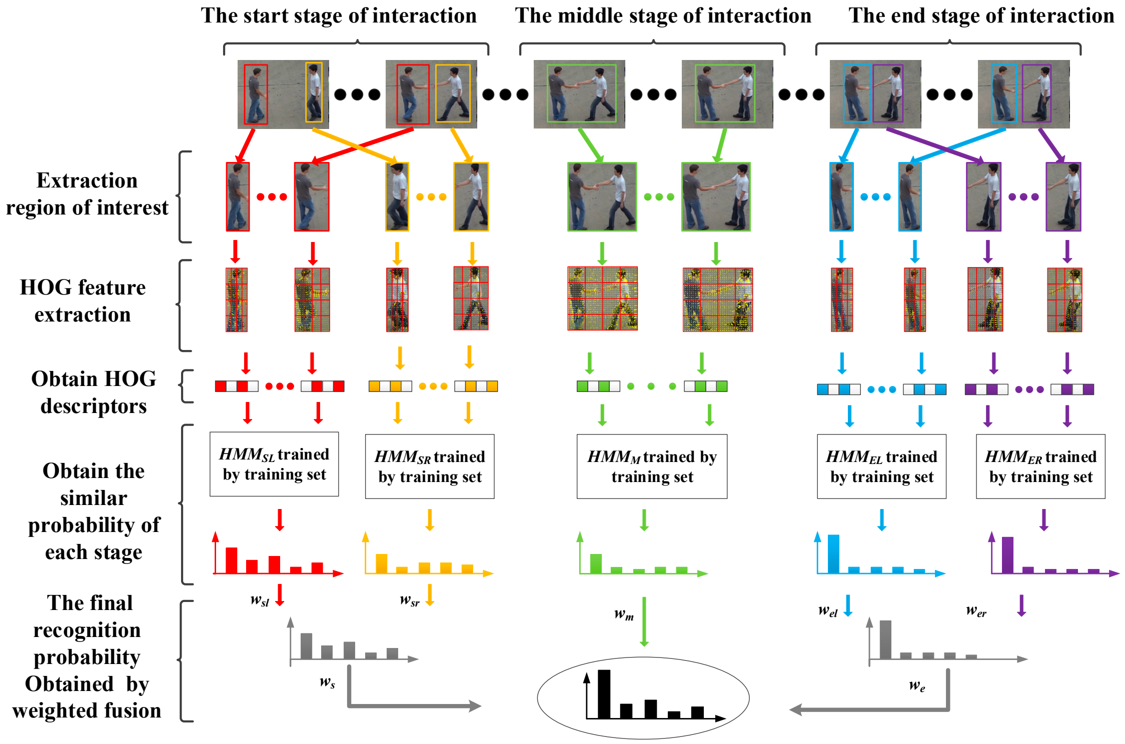

The start stage of interaction
The middle stage of interaction
The end stage of interaction
Extraction region of interest
HOG feature extraction
Obtain HOG descriptors
Obtain the similar probability of each stage
The final recognition probability Obtained by weighted fusion
HMM_{SL} trained by training set
HMM_{SR} trained by training set
HMM_{M} trained by training set
HMM_{EL} trained by training set
HMM_{ER} trained by training set
w_{sl}
w_{sr}
w_{m}
w_{el}
w_{er}
w_{s}
w_{e}

Figure 1. The framework of the proposed recognition method.

2. Piecewise Segmentation of Interactive Behavior

In order to fully characterize the movement relationship between the two sides in the interactive behavior, and also avoid the effect of occlusion in interaction behavior on single atom action recognition, the piecewise segmentation method of interactive behaviors is used in this paper. At first, interactive behavior is divided into three stages, i.e., start stage, middle stage and end stage. Different methods have been adopted in the segmentation and extraction of the region of interest for each stage.

(1) **The start stage of the interactive behavior.** In this stage, the distance between the two persons in the interactive behavior progresses from far to near. The silhouette information of these two persons can be obtained by frame difference method. According to the boundary information, the regions of interaction are obtained with less redundant information. The processing of the progress is shown in Figure 2.

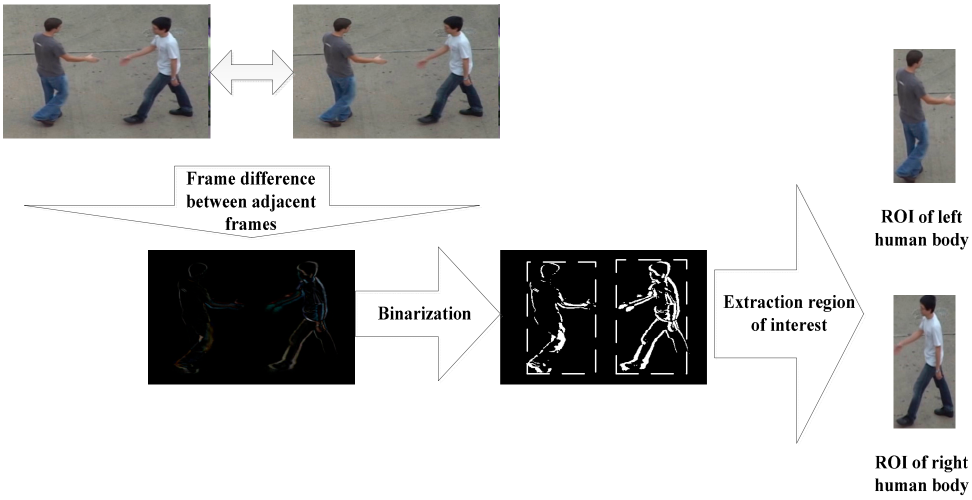

Frame difference between adjacent frames
Binarization
Extraction region of interest
ROI of left human body
ROI of right human body

Figure 2. The region of interest extraction progress in action start stage.

(2) **The middle stage of the interactive behavior.** In this stage, the two persons in the interactive behavior have physical contact. In order to avoid the effect of occlusion in interactive behavior on single atom action recognition, the region of these two persons' bodies is obtained as a whole. The silhouette information of this region can be obtained by frame difference method. The processing of the progress is shown in Figure 3.

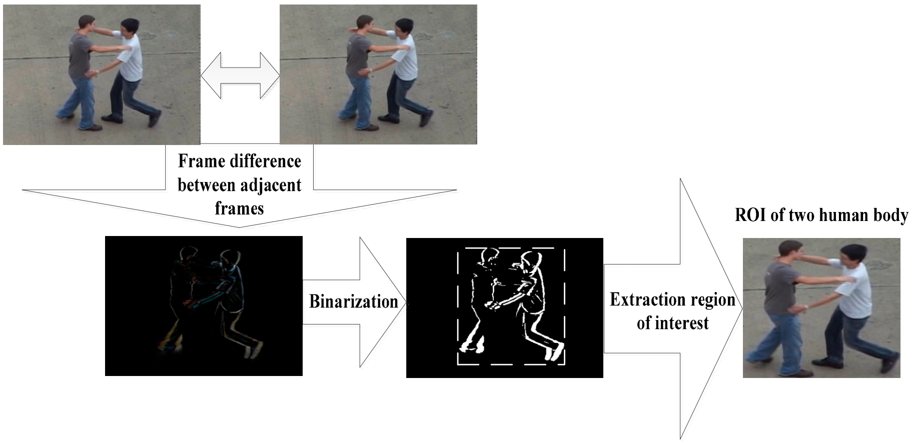

Frame difference between adjacent frames
Binarization
Extraction region of interest
ROI of two human body

Figure 3. The region of interest extraction in action middle stage.

(3) **The end stage of the interactive behavior.** In contrast with the start stage, in this stage, the distance between the two persons in the interactive behavior moves from near to far. The atom actions of these two persons still contain useful information about the interactive behavior. The same method as was used in the start stage is applied to get the region of interest of the two persons in this stage.

In the above three stages, the foreground areas are obtained by performing frame difference. There usually exists some spots of noise and holes inside the moving object, which affects the results of the region of interest. We obtain the ideal region of interest by a series of denoising and morphological operators, the process of which is shown in Figure 4. In a more complex application scenario, we can utilize the Visual background extractor (ViBe) [20] to improve detection accuracy.

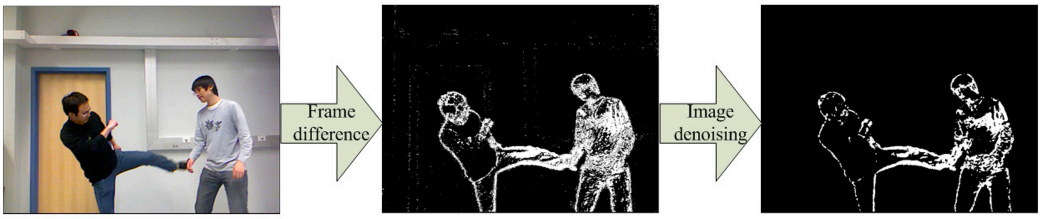


Figure 4. The example of image pre-processing.

3. The Global Feature Extraction and Representation

In order to realize fast and efficient recognition, it is necessary to extract a small amount of raw feature data with a simple and discriminative feature representation. The HOG (Histogram of Oriented Gradient) descriptors was used to represent the ROI of each frame in this paper, since it has been proved that grid-based HOG descriptors significantly outperform existing feature sets for human detection in a previous study [21]. The HOG descriptor [22] reflects the edge gradient information of human motion, does not need a complex edge detection process and is formed by calculating the gradient histogram in local areas. The process of extraction features is shown in Figure 5.

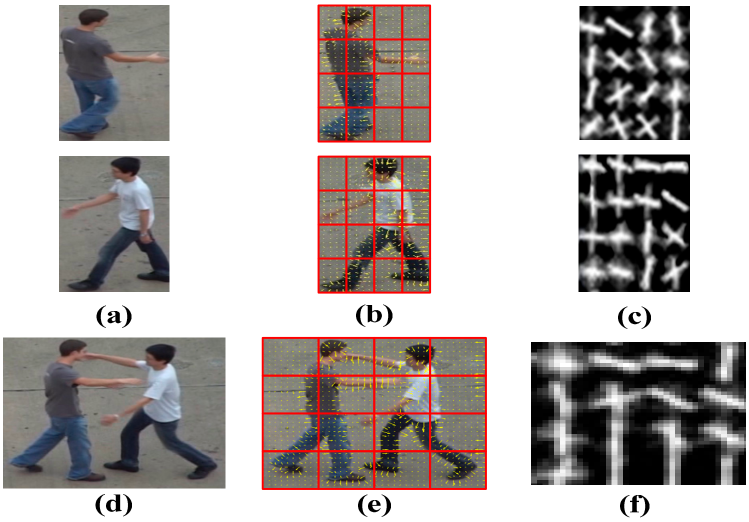


Figure 5. Histogram of Oriented Gradient (HOG) descriptor extraction of region of interest (ROI) in different stages (**a**) original image; (**b**) gradient distribution; (**c**) HOG feature; (**d**) original image; (**e**) gradient distribution; (**f**) HOG feature.

The extraction process of the HOG feature requires two steps:

(1) Calculation of the pixels gradient:

$$T(x_i, y_i) = \sqrt{P_x(x_i, y_i)^2 + P_y(x_i, y_i)^2} \tag{1}$$

$$\theta = \arctan\frac{P_x(x_i, y_i)}{P_y(x_i, y_i)} \tag{2}$$

$T(x_i, y_i)$ represents the amplitude of the gradient, and θ represents the direction of the gradient. $P_x(x_i, y_i) = f(x+1, y) - f(x-1, y)$ and $P_y(x_i, y_i) = f(x, y+1) - f(x, y-1)$ are used to represent and calculate the horizontal gradient and vertical gradient respectively.

(2) Count the histogram of gradient:

The pixels gradient distribution is shown as Figure 5b,e. The gradient distribution is then divided into $ns \times ns$ cells. Finally, the gradient over all the pixels within each cell is projected on m orientations to form a m-dimensions feature vector $\mathbf{v}$. The vector $\mathbf{v}$ should be normalized as Equation (3):

$$\mathbf{v} \rightarrow \frac{\mathbf{v}}{\sqrt{\|\mathbf{v}\|_2^2 + \varepsilon^2}} \tag{3}$$

ε is a small constant.

All the histograms can be concatenated to form an $ns \times ns \times m$ feature vector. In our experiments, we also experimented with ns = 4, 6, 8, 10 and m = 9, 12, 15, 18. The results have shown that the feature reaches the best performance when ns = 4 and m = 12. So the parameters ns = 4 and m = 12 are chosen for our experiment.

4. Piecewise Fusion Recognition Algorithm

4.1. Recognition Based Hidden Markov Model

Hidden Markov model (HMM) can be used to model the motion of small changes in time and space scales, so it is the most widely used method for human action recognition. In order to avoid the problem that the HMM's recognition results could be considered poor because of an incomplete training sample, the proposed method will preserve the probability similarity statistics of the test image sequences at various stages. The final recognition result is obtained by weighted fusion of the similar probability of each stage. The calculation process of the probability similarity statistical graph is shown in Figure 6.

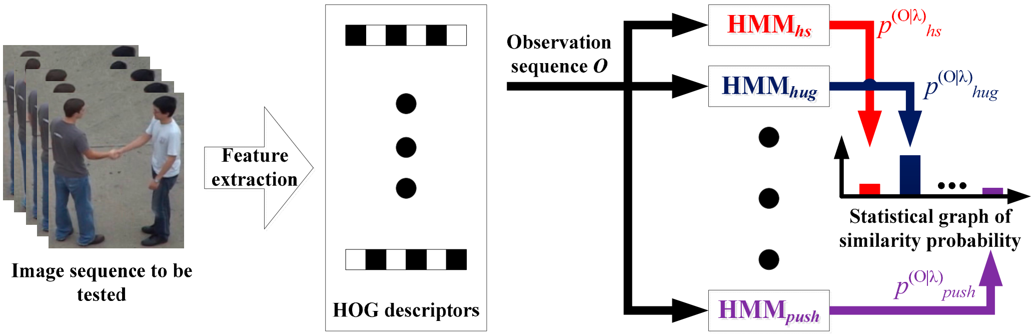


Figure 6. The calculation process of the probability similarity statistics graph with the Hidden Markov model (HMM) algorithm.

The aim of the HMM recognition algorithm for action recognition is to determine the topological structure of the graph model and the calculation of the probability similarity and the training problem of setting the optimal parameters of the model [23]. On the basis of the construction of the HMM topology, the full connected topology of the continuous HMM is utilized to model the action. The probability of generating an observation symbol from each state can be computed by the Gaussian probability-density function as Equation (4):

$$b_i(o_t) = b_{(u_i,\sum_i)}(o_t) = \frac{1}{\sqrt{2p}^d \sqrt{|\sum_i|}} e^{-\frac{1}{2}(o_t-u_i)^T \sum_i^{-1}(o_t-u_i)} \tag{4}$$

where u_i, $\sum_i$ is respectively the mean and covariance matrix of observations in cluster i; d is the dimension of observation symbol o_t; $(o_t - u_i)^T$ is the transpose of matrix $(o_t - u_i)$; $\sum_i^{-1}$ is the inverse of matrix $\sum_i$.

In our method, one interaction sequence is segmented into three stages in time domain. At the start stage and the end stage, the two persons involved in the interaction are segmented into two separate objects, so two HMMs are trained to model the objects' motions in those stages. In the middle stage, the two persons involved in the interaction are segmented as an integrated object, so only one HMM is trained to model the object's motion in this stage. In general, one class of interaction can be represented by five HMMs.

The essence of the HMMs training problem with the given structure is to maximize the observation probability by adjusting the model parameter for the observation sequence. The Baum–Welch algorithm could easily allow us to optimally adapt model parameters to observe the training data [24,25]. However, this is dependent on the choice of initial parameters. If the improper initial parameters are chosen, it can lead the procedure to the Local minimum so that the best action model cannot be obtained. Thus, the result of the *k*-means algorithm is taken as the initial input of the Baum–Welch algorithm.

4.2. Weighted Fusion of Three Stages of Probability Similarity

By using the HMM recognition algorithm, we can obtain the similar probability of the single atom action in the start stage of the interactive behavior, the similar probability of the whole middle stage of the interactive behavior, and the similar probability of the single atom action in the end of the interactive behavior.

The process of the fusion of the three stages of similar probability is divided into the following stages:

(1) **The probability similarity calculation in the start stage of the interactive behavior.** The similar probabilities of the two persons in the start stage of interactive behavior are obtained by using the HMM recognition algorithm. The similar probability of the start stage of interactive behavior can be obtained by weighted fusing of those two similar probabilities. As shown in Equation (5):

$$P_{start} = w_{sl} \times P(O|\lambda)_{sl} + w_{sr} \times P(O|\lambda)_{sr} \tag{5}$$

where P_{start} is the final probability similarity of the start stage of the test image sequence; $P(O|\lambda)_{sl}$ is the probability similarity of the behavior of the left person in the start stage; $P(O|\lambda)_{sr}$ is the probability similarity of the behavior of the right person in the start stage, w_{sl} and w_{sr} are their weights in the weighted fusion process.

(2) **The similarity probability of single atom action in the end stage of the interactive behavior.** The method of weighted fusion in the end stage of the interactive behavior is the same as the method in the start stage, as shown in Equation (6):

$$P_{end} = w_{el} \times P(O|\lambda)_{el} + w_{er} \times P(O|\lambda)_{er} \tag{6}$$

where P_{end} is the final probability similarity of the end stage of the test image sequence; weights w_{el} and w_{er} are obtained by a series of experiments.

(3) **The probability similarity of three stages.** The final recognition probability of the test image sequence is obtained by the weighted fusing of the probability similarity of three stages, as shown in Equation (7):

$$P_{final} = w_s \times P_{start} + w_m \times P_{middle} + w_e \times P_{end} \tag{7}$$

where P_{final} is the final probability similarity of the test image sequence; P_{middle} is the probability similarity of the middle stage of the test image sequence; weights w_s, w_m and w_e are obtained by comparing the average recognition rate of the three stages.

5. Algorithm Verification and Results Analysis

5.1. Algorithm Tested in UT-Interaction Dataset

To test the effectiveness of the proposed method, the UT-Interaction Set 1 benchmark dataset [26] was chosen, which contains six classes of human interactive behaviors performed by 15 people. Each class contains 10 video sequences. This dataset included some challenging factors, such as moving background, cluttered scenes, camera jitters/zooms and different clothes. The experimental database of this paper is composed of all samples of five actions, which were the shake-hand, the hug, the kick, the punch and the push. (The action 'point' was performed by only one person, and therefore not included in the database of this experiment). The actions in the database are shown in Figure 7.

The leave-one-out cross validation method was adopted throughout the process. This involved taking out one action from each action class as test samples, in turns, and then the rest of all the actions as a training set were applied to train HMMs parameters. The circulation continued until all of the actions had completed testing.

The recognition rate was calculated by computing the average of all the circulations.

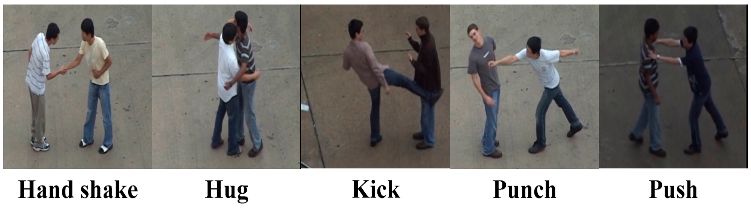


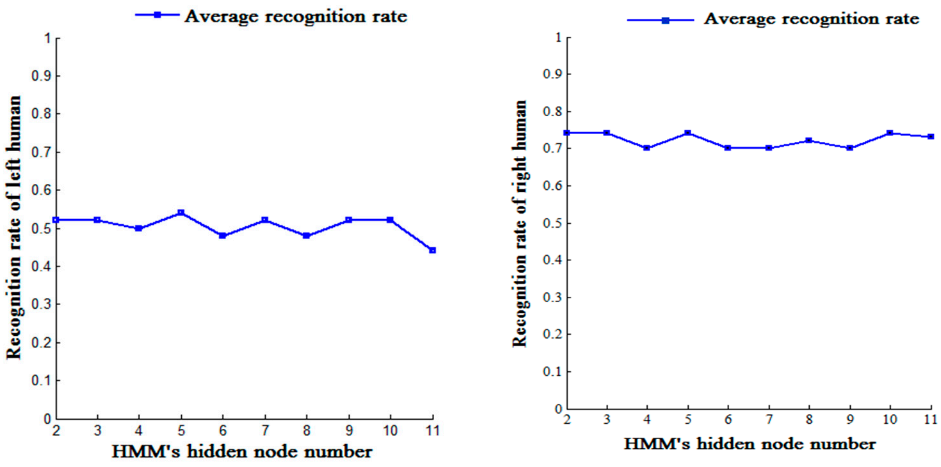
Figure 7. Exemplar frames from UT-interaction dataset.

5.1.1. HMM Number of Hidden Nodes Test in Different Stages

The number of hidden states in the single atom action or interaction of two persons at different stages of interactive behavior has a great influence on the stability of the HMM and the recognition result. The HMM of each stage is trained and tested, and the results are shown as follows:

(1) HMM in the start stage of interactive behavior:

When the ROI is on the left side, the test results of the HMMs' hidden node is shown in Figure 8a. The average recognition rate of the left side of human behavior in the start stage of interactive behavior has reached the highest value when the number of hidden nodes is set to 5. When the ROI is on the right side, the test result of the HMMs' hidden node is shown in Figure 8b. The average recognition rate of the right side human behavior in the start stage of interactive behavior has reached the highest value when the number of hidden nodes is set to 3, 5 and 10. As there is a mutual relationship between the two sides of the interactive behavior, the hidden state of the left side of the human atom behavior in the start stage of the interactive behavior should be corresponding to the hidden state of the right side of human atom behavior, so the hidden node number of the HMM of the right side of human atom behavior should be consistent with the left side. The hidden node number of the HMM of the right side human atom behavior is set to 5 in the final test system.

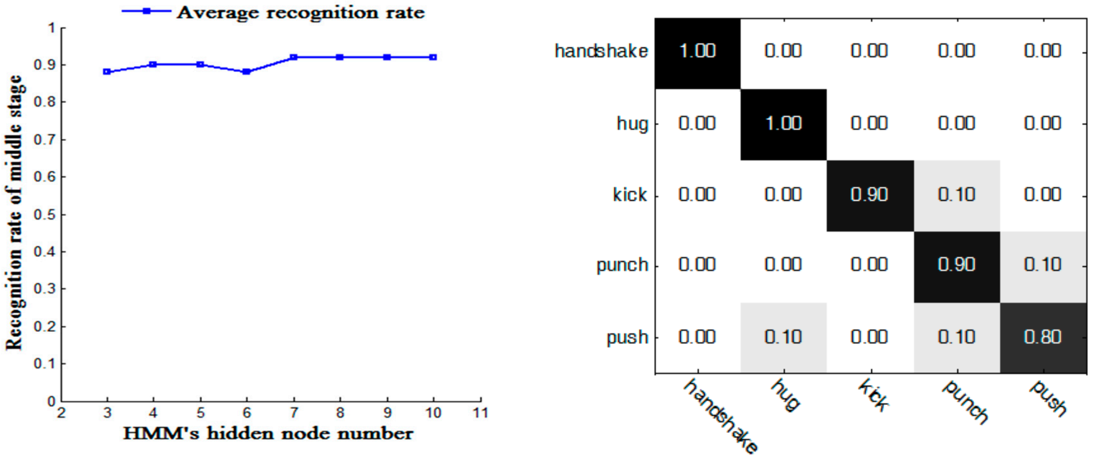


(**a**) (**b**)

Figure 8. The relationship between the average recognition rate and the number of the HMM's hidden nodes in the start stage of the interaction behavior. (**a**) the left side; (**b**) the right side.

(2) HMM in the middle stage of interactive behavior:

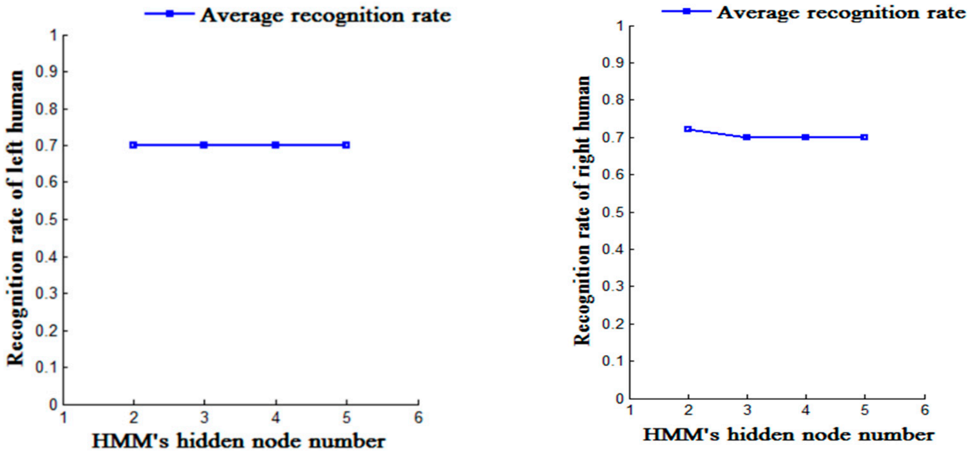
In the middle stage of interactive behavior, the interaction feature is extracted as a whole. The interaction in this stage is more complex than the other stages, so the hidden state of the interactive behavior is more so than in other stages. The test results of the HMM's hidden node in this stage are shown in Figure 9a. The confusion matrix with the optimal number of nodes is shown in Figure 9b.



(**a**) (**b**)

Figure 9. (**a**) The relationship between the number of hidden nodes of the HMM and the average recognition rate, (**b**) the confusion matrix of the optimal results in the middle stage HMM in the end stage of interactive behavior.

The information of the interactive behavior in this stage is always less than the other stages, as is the case with the number of hidden states of the HMM in this stage. The test result of the HMM's hidden node number of the left side of the interaction in the end stage is shown as Figure 10a. The test result of the HMM's hidden node number of the right side of the interaction in the end stage is shown as Figure 10b. As the end stage contains less information, the number of hidden states is less than other stages, which is what was predicted in this paper. Combined with the test results in Figure 10b, the hidden node number of the HMM of the left and right side human atom behavior is set to 2 in the final test system.

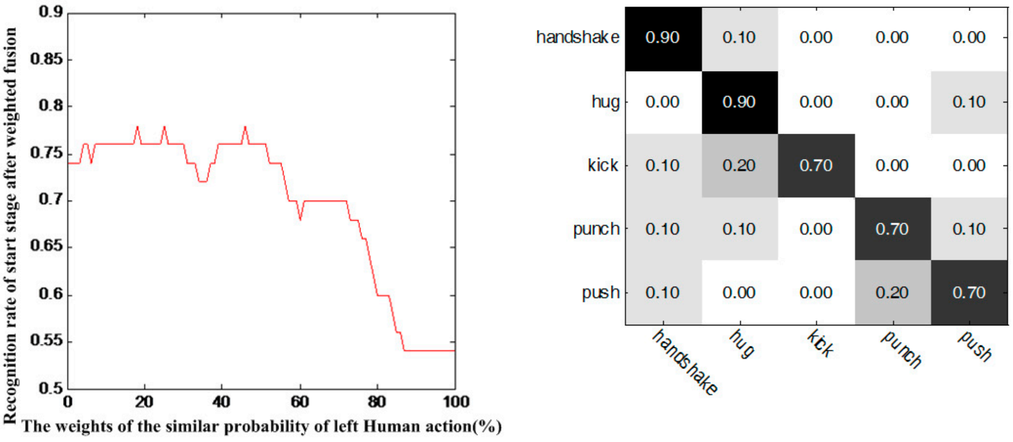


(**a**) (**b**)

Figure 10. The relationship between the average recognition rate and the number of the HMM's hidden nodes in the end stage of the interaction behavior, (**a**) the left side, (**b**) the right.

5.1.2. Weighted Fusion Test

(1) **At the start stage of the interaction,** the regions of the left and right sides of the moving target are extracted respectively, and the regions of interest are described by using HOG descriptors. The probability similarity of them are obtained by using a HMM which contain five hidden nodes. The final recognition results are obtained by weighted fusion of their probabilities similarity. The optimal weights are obtained by a large number of experiments, as shown in Figure 11a. The optimal weights of weighted fusion are 18% and 82% respectively. The final confusion matrix of recognition results in the start stage is as shown in Figure 11b.



	handshake	hug	kick	punch	push
handshake	0.90	0.10	0.00	0.00	0.00
hug	0.00	0.90	0.00	0.00	0.10
kick	0.10	0.20	0.70	0.00	0.00
punch	0.10	0.10	0.00	0.70	0.10
push	0.10	0.00	0.00	0.20	0.70

(**a**) (**b**)

Figure 11. (**a**) Weighted fusion optimal parameters test, (**b**) confusion matrix of recognition results after the weighted fusion of the start stage (Average recognition rate: 78%).

At the start stage of interactive behavior, the characteristics of the action are often not obvious, and some actions at the start stage are very similar, so the recognition result of the single atom action at the start stage is not good. However, the performance of recognition has been greatly improved by weighted fusing. Experimental results are shown in Table 1.

Table 1. Recognition results of single atom behavior and weighted fusion recognition results of start stage.

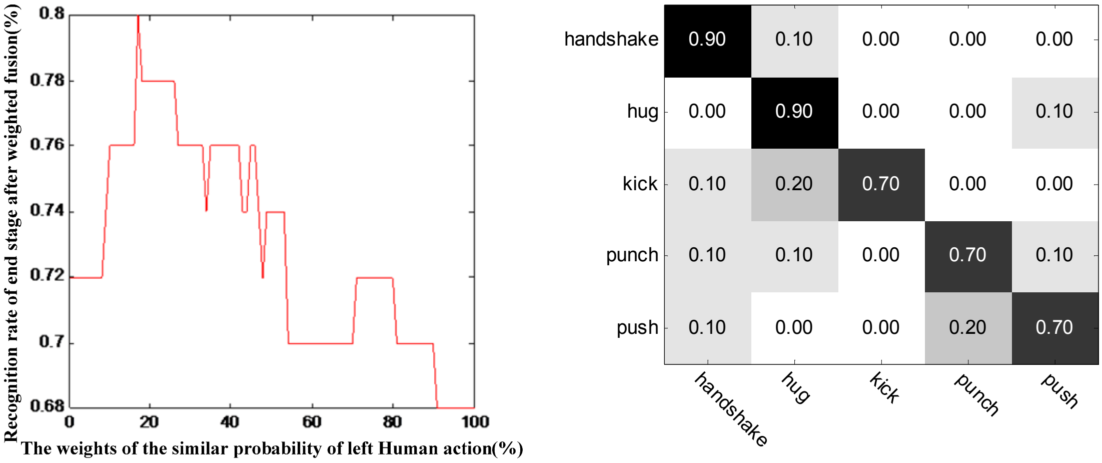
Action Type	Hand Shake	Hug	Kick	Punch	Push	Avr/%
Left	80	40	50	50	50	54
Right	90	90	60	60	70	74
Final	90	90	70	70	70	78

(2) **At the end stage of the interaction,** the regions of the left and right sides of the moving target are extracted respectively, and the regions of interest are described by using HOG descriptors. The probability similarity of them are obtained by using a HMM which contains two hidden nodes. The final recognition results are obtained by weighted fusion of their similar probabilities. As was the case with the start stage, the recognition results of the single atom behavior are not good, but after the weighted fusion, the performance of recognition has been greatly improved. Experimental results are shown in Table 2.

Table 2. Recognition results of single atom behavior and weighted fusion recognition results of end stage.

Action Type	Hand Shake	Hug	Kick	Punch	Push	Avr/%
Left	100	80	40	50	70	68
Right	70	70	90	50	80	72
Final	90	70	90	70	80	80

The optimal weights are obtained by a large number of experiments, as shown in Figure 12a, the optimal weights of weighted fusion are 17% and 83% respectively. The final confusion matrix of the recognition results after the weighted fusion of the start stage is shown in Figure 12b.



	handshake	hug	kick	punch	push
handshake	0.90	0.10	0.00	0.00	0.00
hug	0.00	0.90	0.00	0.00	0.10
kick	0.10	0.20	0.70	0.00	0.00
punch	0.10	0.10	0.00	0.70	0.10
push	0.10	0.00	0.00	0.20	0.70

(**a**) (**b**)

Figure 12. (**a**) Weighted fusion optimal parameters test, (**b**) confusion matrix of recognition result after weighted fusion of the start stage (Average recognition rate: 80%).

(3) **Weighted fusion of similar probabilities of the three stages.** According to the weighted fusion process of the start stage and the end stage, it can be found that the average recognition rate is related to the optimal weights in the weighted fusion of the different stages of the interaction behavior. In the process of weighted fusion, the weights are set according to the recognition results of the different stages in this paper. The weight of the similar probability at the start stage is 20% (the average recognition rate of this stage is 78%). The weight of the similar probability of the middle stage is 48% (the average recognition rate of this stage is 92%). The weight of the similar probability of the middle stage is 32% (the average recognition rate of this stage is 80%).

Experimental results are shown in Table 3. The recognition rate of the actions 'handshake' and 'hug' obtained by using weighted fusion method reached 100%. The recognition rate of the actions 'kick' 'punch' and 'push' obtained by using weighted fusion method reached 90%. The recognition result obtained by using the weighted fusion method is better than the recognition results of all stages. The confusion matrix is shown in Figure 13.

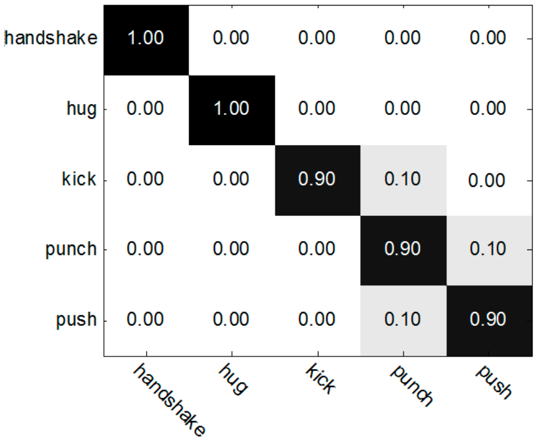


Figure 13. Confusion matrix of the recognition results obtained by using the weighted confusion method.

The method proposed in this paper accurately recognizes most of the interactive behavior, with the average recognition rate obtained by using this method being 94%, and in particular, the recognition rate of the actions of 'hand shake' and 'hug', which reached 100%. Through the observation of the confusion matrix and database sample, the observation angle of view leads to the high similarity between action 'kick', 'push' and 'boxing', and provides a reference for further improvement of the method.

5.1.3. The Comparison of the Performance

The comparisons of performance between the proposed method and related recent works based on the UT-Interaction dataset are shown in Table 3.

Table 3. Comparison with related work in recent years.

Source	Year	Method	Acc/%
Our approach	2017	Stage model + HOG + HMM	94
Kong et al. [16]	2014	global template + local 3D feature + discriminative model	85
Mukherjee et al. [27]	2011	Bipartite graph + key pose doublets	79.17
Brendel et al. [28]	2011	tubes + spatio-temporal relationships graph model	78.9
Liang et al. [29]	2016	Spatio-temporal features_context	92.3

Obviously, our approach has achieved the best recognition result. Compared with other recognition methods, the advantage of the proposed method in this paper is that the feature extraction is very simple with a quick process speed.

5.2. Algorithm Tested in SBU-Interaction Dataset

The SBU-interaction two-person interaction action video dataset is used to test our proposed method. The dataset is a public dataset, and makes use of the Microsoft Kinect sensor to create a two-person interaction action with depth of image, color images and frame images. The dataset includes eight types of interaction actions (approaching, departing, kicking, punching, pushing, hugging, shaking hands, exchanging). A total of seven people are involved in action-taking in the same laboratory environment. Each action is performed by different people, and the entire dataset has 280 sets of interactive actions [13]. The examples of the dataset are shown in Figure 14.

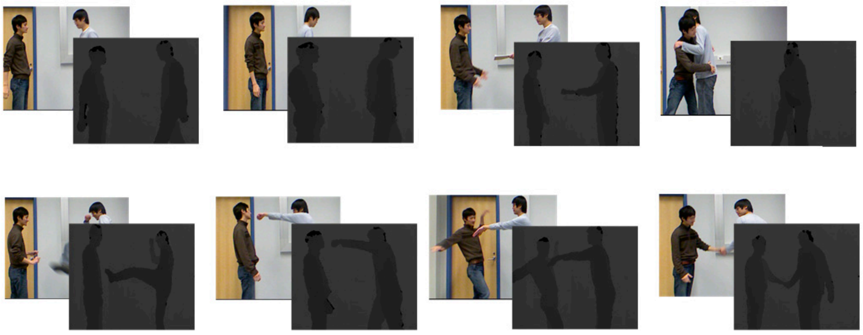


Figure 14. The examples of the dataset.

5.2.1. Experimental Results and Analysis

Taking into account the large number of videos in the dataset, we randomly selected a certain amount of test data from the dataset, and selected the other data as training data. Two experiments were tested in this dataset:

(1) Interaction recognition by weighted fusion of probability of the RGB and depth videos: In this experiment, piecewise segmentation of the interactive behavior was not performed. The HOG features were extracted in RGB and depth image respectively. The parameters in feature extraction were chosen to be the same as in the above experiment. Only two HMMs were trained to model one human interaction with features extracted in RGB video and features extracted in depth video. The final recognition result of human interaction was obtained by weighted fusion of the probabilities of the two datasets. The weighted parameters of RGB video and depth video probability were 40% and 60%. The results are shown in Table 4.

(2) Interaction recognition by weighted fusion of probability of the RGB and depth videos and three stages: In this experiment, piecewise segmentation of interactive behavior was performed. The HOG features were extracted in RGB and depth image respectively. The parameters in feature extraction and weight parameters were chosen to be the same as in the above experiment. Ten HMMs were trained to model one human interaction with features extracted in RGB video and features extracted in depth video and three stages. The final recognition result of human interaction was obtained by weighted fusion of the probabilities of all probabilities. The results are shown in Table 4.

Table 4. Recognition results comparison of above two experiments.

Methods	Without Piece Fusion	With Piece Fusion
RGB image	74.86%	80.00%
Depth image	78.38%	86.67%
Weighted fusion	85.88%	91.70%

As can be seen from Table 4, the recognition accuracy of the weighted fusion of two-source information is much better than that of single-source information. Furthermore, compared with the results without piece fusion, using piece fusion can greatly improve the accuracy of two-person interactive action recognition.

5.2.2. The Comparison of the Performance

The comparison of recognition rates between the proposed method and the recent methods based on SBU-interaction two-person interactive action dataset is shown in Table 5. It can be seen that the skeleton structure model for the two-person interactive action recognition is used by the authors of [13,30]. Obviously the recognition rate of the proposed method in this paper is superior to that of the results in study [8,15] The proposed method does not need human skeleton information. The features are directly extracted from the depth images. The extraction method is simple and easy to implement.

Table 5. Comparison with related work in recent years.

Comparative Literature	Features and Recognition Methods	Recognition Rate (%)
Yun et al. [13]	Joint distance + SVM Joint distance + MILBoost	87.6% 91.1%
Yanli Ji et al. [30]	BOW CFDM	82.5% 89.4%
Our paper method	HOG(RGB) + HOG(depth) without piece weighted fusion HOG(RGB) + HOG(depth) + piece weighted fusion	85.9% 91.7%

6. Conclusions

A novel interactive behavior recognition method based on multiple stage probability fusion is proposed in this paper. In order to preserve the action characteristics of the interaction, and reduce the redundant background information of the complex environment, the regions of the two persons are segmented and extracted respectively at the start stage and the end stage of interactive behavior. In order to avoid the segmentation error caused by human occlusion, the region of these two people is segmented and extracted as a whole at the middle stage of interactive behavior. The probability similarity of each stage is obtained by using HMMs, and the final recognition result is obtained by weighted fusion of those similar probabilities. A large number of experiments on the UT-interaction database and the SBU-interaction dataset demonstrated that the method is simple and has better recognition ability of the interaction behavior.

Acknowledgments: This work is supported by the Program for Liaoning Excellent Talents in University (LR2015034) and Natural Science Foundation of Liaoning Province (201602557).

Author Contributions: X.J. and Z.J. conceived and designed the experiments; X.J. and C.W. performed the experiments and analyzed the data; X.J. and Z.J. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, Segmentation and Recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
2. Ji, X.; Liu, H. Advances in View-Invariant Human Motion Analysis: A Review. *IEEE Trans. Syst. Man Cybern. C* **2010**, *40*, 13–24.
3. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust Human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2016**, *61*, 295–308. [CrossRef]
4. Jalal, A.; Kamal, S.; Kim, D. Depth Silhouettes Context: A New Robust Feature for Human Tracking and Activity Recognition Based on Embedded HMMs. In Proceedings of the International Conference on Ubiquitous Robots and Ambient Intelligence, Goyang, Korea, 28–30 October 2015; pp. 1–7.
5. Farooq, A.; Jalal, A.; Kamal, S. Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. *KSII Trans. Internet Inf. Syst.* **2015**, *9*, 1856–1867.
6. Jalal, A.; Kamal, S.; Kim, D. Human depth sensors-based activity recognition using spatiotemporal features and hidden markov model for smart environments. *J. Comput. Netw. Commun.* **2016**, *2016*, 8087545. [CrossRef]
7. Gaur, U.; Zhu, Y.; Song, B. A "String of Feature Graphs" Model for Recognition of Complex Activities in Natural Videos. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2595–2602.

8. Yu, T.; Kim, T.; Cipolla, R. Real-time Action Recognition by Spatio-Temporal Semantic and Structural Forests. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 31 August–3 September 2010; pp. 1–12.
9. Burghouts, G.J.; Schutte, K. Spatio-temporal layout of human actions for improved bag-of-words action detection. *Pattern Recognit. Lett.* **2013**, *34*, 1861–1869. [CrossRef]
10. Peng, X.; Peng, Q.; Qiao, Y. Exploring Dense Trajectory Feature and Encoding Methods for Human Interaction Recognition. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Huangshan, China, 17–19 August 2013; pp. 23–27.
11. Li, N.; Cheng, X.; Guo, H.; Wu, Z. A Hybrid Method for Human Interaction Recognition Using Spatio-Temporal Interest Points. In Proceedings of the International Conference on 22nd Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 2513–2518.
12. Bingbing, N.; Guang, W.; Pierre, M. RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–11 November 2011; pp. 1147–1153.
13. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, D. Two-Person Interaction Detection Using Body-Pose Features and Multiple Instance Learning. In Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 1–8.
14. Patron-Perez, A.; Marszalek, M.; Reid, I.; Zisserman, A. Structured learning of human interactions in TV shows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2441–2453. [CrossRef] [PubMed]
15. Raptis, M.; Sigal, L. Poselet Key-framing: A Model for Human Activity Recognition. In Proceedings of the IEEE conference on Computer Vision Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2650–2657.
16. Kong, Y.; Jia, Y.; Fu, Y. Interactive Phrases: Semantic descriptions for human interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1775–1788. [CrossRef] [PubMed]
17. Slimani, K.; Benezeth, Y.; Souami, F. Human Interaction Recognition Based on the Co-Occurrence of Visual Words. In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 461–466.
18. Dong, Z.; Kong, Y.; Liu, C.; Li, H.; Jia, Y. Recognizing Human Interaction by Multiple Features. In Proceedings of the First Asian Conference on Pattern Recognition, Beijing, China, 28 November 2011; pp. 77–81.
19. Kong, Y.; Liang, W.; Dong, Z.; Jia, Y. Recognising human interaction from videos by a discriminative model. *IET Comput. Vis.* **2014**, *8*, 277–286. [CrossRef]
20. Barnich, O.; Van Droogenbroeck, M. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Trans. Image Process.* **2011**, *20*, 1709–1724. [CrossRef] [PubMed]
21. Dalad, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
22. Ji, X.; Zhou, L.; Li, Y. Human Action Recognition Based on AdaBoost Algorithm for Feature Extraction. In Proceedings of the IEEE Conference on Computer and Information Technology, Xi'an, China, 22–23 December 2014; pp. 801–805.
23. Rabiner, L.R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 267–296. [CrossRef]
24. Jalal, A.; Uddin, M.; Kim, T. Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Trans. Consum. Electron.* **2012**, *58*, 863–871. [CrossRef]
25. Jalal, A.; Kamal, S.; Kim, D. A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. *Sensors* **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
26. Ryoo, M.S.; Aggarwal, J.K. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In Proceedings of the IEEE Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1593–1600.
27. Mukherjee, S.; Biswas, S.K.; Mukherjee, D.P. Recognizing Interaction between Human Performers Using 'Key Pose Doublet'. In Proceedings of the 19th ACM Multimedia Conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 1329–1332.

28. Brendel, W.; Todorovic, S. Learning Spatiotemporal Graphs of Human Activities. In Proceedings of the IEEE Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 778–785.
29. Liang, J.; Xu, C.; Feng, Z.; Ma, X. Affective interaction recognition using spatio-temporal features and context. *Comput. Vis. Image Underst.* **2016**, *144*, 155–165. [CrossRef]
30. Ji, Y.; Cheng, H.; Zheng, Y.; Li, H. Learning contrastive feature distribution model for interaction recognition. *Vis. Commun. Image Represent.* **2015**, *33*, 340–349. [CrossRef]