sensors

MDPI

Article

Human Interaction Recognition Based on Whole-Individual Detection

Qing Ye *, Haoxin Zhong, Chang Qu and Yongmei Zhang

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; frank097u7@163.com (H.Z.); qc_qearl@163.com (C.Q.); zhangym@ncut.edu.cn (Y.Z.)

* Correspondence: yeqing@ncut.edu.cn; Tel.: +86-13520155178

Received: 12 March 2020; Accepted: 18 April 2020; Published: 20 April 2020

Abstract: Human interaction recognition technology is a hot topic in the field of computer vision, and its application prospects are very extensive. At present, there are many difficulties in human interaction recognition such as the spatial complexity of human interaction, the differences in action characteristics at different time periods, and the complexity of interactive action features. The existence of these problems restricts the improvement of recognition accuracy. To investigate the differences in the action characteristics at different time periods, we propose an improved fusion time-phase feature of the Gaussian model to obtain video keyframes and remove the influence of a large amount of redundant information. Regarding the complexity of interactive action features, we propose a multi-feature fusion network algorithm based on parallel Inception and ResNet. This multi-feature fusion network not only reduces the network parameter quantity, but also improves the network performance; it alleviates the network degradation caused by the increase in network depth and obtains higher classification accuracy. For the spatial complexity of human interaction, we combined the whole video features with the individual video features, making full use of the feature information of the interactive video. A human interaction recognition algorithm based on whole–individual detection is proposed, where the whole video contains the global features of both sides of action, and the individual video contains the individual detail features of a single person. Making full use of the feature information of the whole video and individual videos is the main contribution of this paper to the field of human interaction recognition and the experimental results in the UT dataset (UT–interaction dataset) showed that the accuracy of this method was 91.7%.

Keywords: human interaction recognition; whole-individual detection; parallel multi-feature fusion network; Gaussian model downsampling

1. Introduction

Human motion recognition is a key technology for intelligent video surveillance. It is widely used in various scenarios such as human–computer interaction [1], motion analysis [2–7], intelligent monitoring, gesture recognition [8,9], and facial emotion recognition [10–13]. Human motion recognition is divided into single-person motion recognition and multi-person interactive motion recognition. At present, the research on human interaction recognition has gradually attracted the attention of researchers, and has obtained certain research results [14–20]. The algorithm framework of human interaction recognition is generally divided into whole human interaction recognition and individual human interaction recognition. This paper combined the characteristics of the two ways, and proposed a method of human interaction recognition based on whole and individual detection.

1.1. Related Work

At present, there are many practical algorithms and networks that can be used for human motion video recognition research. Zhao et al. [21] believe that convolution directly in the time dimension implies a strong assumption that the features between frames are well aligned, but in fact, the person or the object may perform large displacement or deformation in the video. They proposed considering trajectory convolution as a special case of 3D deformable convolution, which provides the offset amount by time series information, so that the trajectory convolution can be easily realized based on the code of the deformable convolution. Silambarasi et al. [22] proposed a video representation method of 3D volume space and human motion trajectory that projected the video onto three different views, called the 3D space–time plane, and used time-view motion tracking to identify various human behaviors. Du Tran et al. [23] proposed a new form of convolution to deal with spatiotemporal information and used 3D convolution to process spatio-temporal information and made further improvements in the form of 3D convolution. They proposed an improved model: 2D + 3D, called mixed convolution, which has the benefit of reducing parameters and maintaining performance. Chen et al. [24] made the convergence speed as fast as possible and proposed a multi-fiber network. This splits the complex network into a lightweight network and uses the information flow between the fibers to introduce the multiplexer module. Wang et al. [25] proposed a non-local operation as a generic family of building blocks for capturing long-range dependencies. This was then a concrete introduction to their method, which was inspired by the classical non-local means method of computer vision.

In the field of human interaction recognition, there are many practical algorithms and networks that can be used in human interactive video recognition research. Nilar Phyo et al. [26] applied deep learning technology over the skeleton motion history image (Skl MHI) [27] of human actions to implement HAR (human action recognition) that can work independently on the problem domain. Li et al. [28] treated the interactive actions as individual motions, combining global features and local features to identify human interactions. The algorithm framework for human interaction recognition is generally divided into whole body-based human interaction classification and individual-based human interaction classification [29]. Among them, the whole recognition method refers to describing the human interaction as a whole, including all the people involved in the interaction in the video. Guo et al. [30] categorized the multi-person interaction as individual layer or interaction layer, and proposed a hidden Markov model based on observation vector decomposition. Xiaofei Ji et al. [31] introduced a hierarchical structure of interactive action recognition based on the process of human interaction. According to chronological order, the actions are divided into action start time, action execution time, and action end time. Vahdat et al. [32] proposed an interactive action recognition algorithm based on chronological key poses by treating two interactions as two individuals, learning the model parameters of each individual, and then identifying them. However, this method could not capture the human interaction information. Such methods mainly deal with individual actions, which may interfere with the action classification results due to the existence of individual occlusion and self-occlusion.

This paper focused on human interaction recognition in video. In the existing video research algorithms for human interaction motion, many methods recognize the two sides of the action as a whole and lose the characteristic information that the individual brings. There are also a few algorithms that separate the two sides of the action into two separate individuals for recognition, but they ignore the characteristic information that the whole brings. Therefore, this paper proposed a human interaction recognition framework based on whole–individual detection. This method contains the characteristics of the whole information and the characteristics of the individual information. The whole video includes both sides of the action, and can extract global features such as the relative position and orientation of the action, and the individual detection video contains a single person, where the individual detailed action feature information can be extracted.

1.2. Contribution

First, we propose an improved fusion time-phase feature of the Gaussian model to obtain video keyframes and remove the influence of a large amount of redundant information. This method resolves the problem of the differences in action characteristics at different time periods.

Then, we propose a multi-feature fusion network algorithm based on parallel Inception and ResNet. This multi-feature fusion network not only reduces the network parameter quantity, but also improves the network performance; it alleviates the network degradation caused by the increase in network depth and obtains higher classification accuracy. This algorithm solves the problem of the complexity of human interaction feature extraction.

We have noted that the whole video contains the global features of both sides of action, and that the individual video contains the individual detail features of a single person. Therefore, we combined the whole video features with the individual video features, making full use of the feature information of the interactive video, which will help us solve the problem of the spatial complexity of human interaction.

2. Proposed Method

Due to the spatial complexity of human interaction, we propose a two-person interaction recognition algorithm based on whole–individual detection. The video includes two individuals of the action. Information can be extracted such as the relative position and orientation of the interactive action, and the individual detailed action features can be collected from the individual action video. As shown in Figure 1, in the motion video individual detection stage, the whole video is split into individual video (left) and individual video (right). This paper used HOG (histogram of oriented gradient) [33] and SVM (support vector machine) [34] and the Kalman tracking algorithm [35] to obtain the position of individuals for video detection as the two algorithms can reduce the impact of occlusion on pedestrian detection. In the feature extraction and model training stage, first, the image is preprocessed by data enhancement and normalization. Then, the video downsampling algorithm based on Gaussian distribution is used to improve the validity of the data. Finally, the parallel multi-feature fusion network is proposed for model training. In the action video recognition stage, the preliminary recognition results that are obtained by the whole video and the individual segmented video are fused. Then, we combine the preliminary recognition results decision level to obtain the final results.

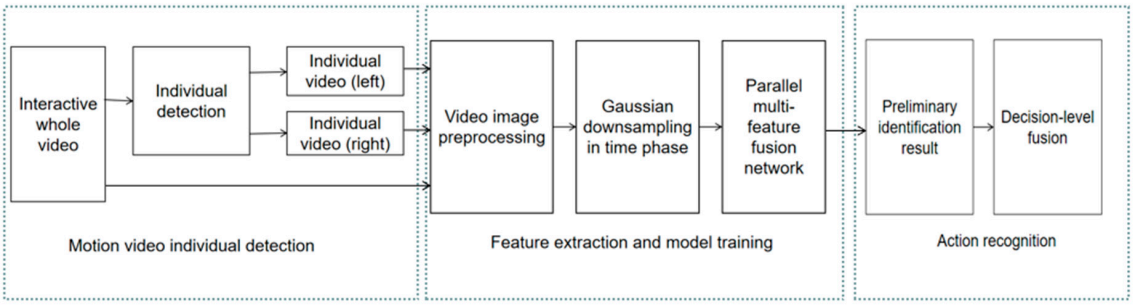

Figure 1. Human interaction recognition based on whole–individual detection.

2.1. Motion Video Individual Detection

Due to the spatial complexity of interaction features, the motion characteristics of individual movements and the characteristics of interaction movements, we proposed an interactive motion recognition framework based on whole–individual detection. In the individual detection part of the interactive action video, the moving target is detected for the action video. The video detection is performed according to the detection result, and the algorithm block diagram used is shown in Figure 2.

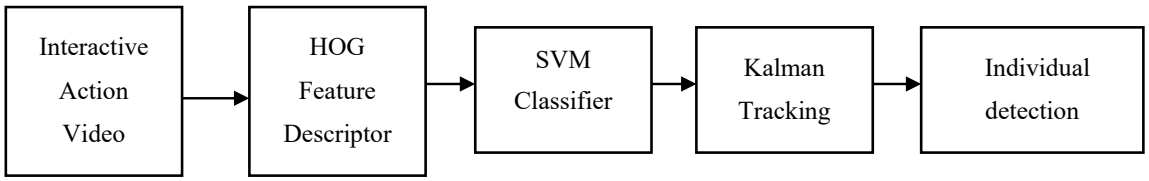

Figure 2. Block diagram of the individual detection algorithm for interactive video.

When the HOG feature descriptor and the pedestrian detection algorithm of the SVM classifier are applied to the interactive action video, loss detection and false detection phenomenon may take place. In this paper, a Kalman filter-based auxiliary tracking model was added to the pedestrian detection process to track each individual in the human interaction video. The trajectory and other information detected by consecutive frames provide reasonable prediction of the target's next frame position. Human body loss detection is greatly reduced by using target tracking. As shown in Figure 3, a complete two-person interactive motion video is segmented into two individual motion videos, recorded as individual video (left) and individual video (right). First, we input the human interactive action video. Then, we extracted HOG features from the video. After that, we imported the parameters into the SVM classifier for training. Finally, we used the Kalman tracking algorithm to achieve individual detection.

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

Figure 3. Video individual detection diagram: (**a**) Original video; (**b**) Individual video (left); (**c**) Individual video (right).

2.2. *Video Downsampling with Time-Phased Features*

When using video images for feature extraction and training, the choice of video sampling methods will directly affect the generalization ability of the classification model. If there is no video downsampling, adjacent video will generate a lot of redundant information, which will increase the burden on the network. Considering the difference in the phase characteristics of the action time, this paper proposed a Gaussian model downsampling method that combines time-phase features.

The Gaussian function is also called a normal distribution, and the expression of the probability density function is as shown in Equation (1).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

where μ is the expected value of the Gaussian distribution and σ is the standard deviation of the Gaussian distribution, which is related to the magnitude of the Gaussian distribution. It can be seen from the distribution curve that the closer to the expected value, the higher the probability density value. There is a general rule in human interaction video: the closer to the trigger stage of the target action, the higher the distinction between actions. We associated this rule with a Gaussian probability density function that is very similar. In the public dataset taken in this paper, each action video sample could be divided into the action start period, action execution period, and action end period. Based on the analysis, it can be seen that the action execution period had significant feature differences, which can help to classify the movements, and the action frame was concentrated in the center. Analogous to the Gaussian model, we proposed a video downsampling method based on Gaussian probability distribution. In the process of video downsampling, we adopted different sampling methods according to the different action stages of the video. For the video from the beginning to the end of the target action, we sampled at small intervals, and for the redundant video outside the target action, we sampled at large intervals. The selection of the sampling interval needs to be obtained through repeated experiments. The video features will also be more differentiated after Gaussian downsampling processing.

2.3. Human Interaction Feature Extraction Based on Parallel Multi-Feature Network

In order to improve the accuracy of interactive action video recognition, a convolutional neural network based on parallel multi-feature fusion was proposed for the extraction of interactive feature information. The migration learning method is adopted in the feature extraction process [33]. A block diagram of the algorithm based on an Inception [36] and ResNet (Residual Network) [37] parallel multi-feature fusion network is shown in Figure 4.

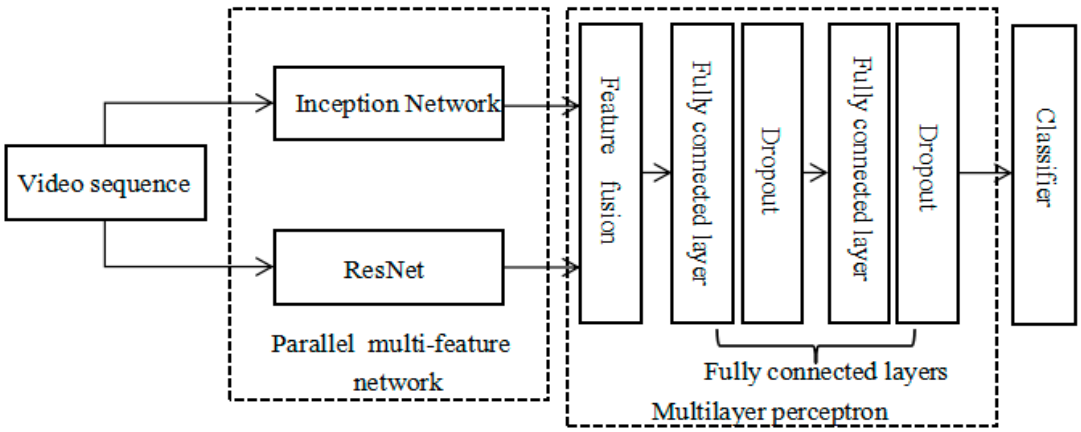

Figure 4. Block diagram of the parallel multi-feature fusion network algorithm.

The Inception V3 [38] network is an important breakthrough in the history of convolutional neural networks. We used the Inception pre-training network to obtain feature information. In general, in order to improve network performance, the most common way for researchers is to increase the depth and width of the network, but this approach will generate a huge number of parameters. As the number of network layers increases, the training process will be more cumbersome and will easily lead to over-fitting of the data. In order to ensure the performance of computing while expanding the

network, the Inception network emerges. As shown in Figure 5, the network structure clusters the sparse matrix into more dense sub-matrices to improve the computing performance. The depth and width of the neural network are modified. The large convolution kernel is split into small convolution kernels of different sizes such as 1×1, 3×3, and 5×5.

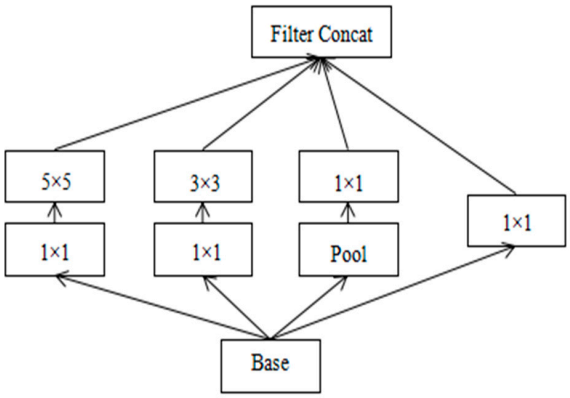

Figure 5. Inception module.

The Inception structure uses different receptive fields to fuse different scale features, and the obtained image feature information is more complete. At the same time, the series of networks eliminates the full connection layer in the network output layer. This method greatly reduces the parameters of the entire network and improves the efficiency of the training process. With the rapid development of deep learning technology, the Inception network is constantly improving and innovating. For example, Inception V2 network [36] introduces the batch normalization layer, and uses two 3×3 convolution kernels to represent a 5×5 convolution kernel. The depth of the neural network improves its nonlinearity. The concept of asymmetric convolution is proposed in the Inception V3 network structure. In this structure, the large convolution kernel is further decomposed. In order to further reduce the number of parameters in the network and improve the speed of the operation, the network further decomposes the $N \times N$ scale convolution kernel and decomposes it into a $1 \times N$ convolution kernel and an $N \times 1$ convolution kernel. It can not only improve the operation speed, but also alleviate the data over-fitting phenomenon.

We continued to explore the feature extraction method based on residual neural network [37]. As the convolutional neural network goes deeper and deeper, a series of problems have emerged. In the process of network training, as the number of network layers deepens, gradient disappearance or gradient explosion may occur. These problems make the network difficult to converge. Researchers hope to increase the nonlinearity of neural networks by increasing the number of neural network layers. At the same time, they are trapped by the phenomenon of network degradation. Figure 6 shows the principle framework of the deep residual network.

In the residual network structure diagram of Figure 6, the input of x is directly transferred to the output as the initial result by ways of "shortcut connections". The output result is $H(x) = F(x) + x$. When $F(x) = 0$, then $H(x) = x$. This is identity mapping (the output is equal to the input). Therefore, ResNet is equivalent to changing the learning goal, which is no longer to learn a complete output, but the difference between the target value $H(x)$ and x, called residual $F(x) = H(x) - x$. Therefore, the following training goal is to approximate the residual result to 0, so that as the network deepens, the accuracy does not decrease.

There is a residual unit that can be expressed in the form of Equations (2) and (3):

$$y_l = h(x_l) + F(x_l, W_l)\left(W_l = \left\{W_{l,k} | 1 \le k \le K\right\}\right) \tag{2}$$

$$x_{l+1} = f(y_l) \tag{3}$$

where x_l, y_l represent the input and output values of the l_{th} neuron; $h(x_l)$ is the representative identity map; and $f(y_l)$ is the representation activation function. Under the condition of identity mapping: $h(x_l) = x_l$, $y_l = f(y_l)$, there is Equation (4).

$$x_{l+1} = x_l + F(x_l, W_l) \tag{4}$$

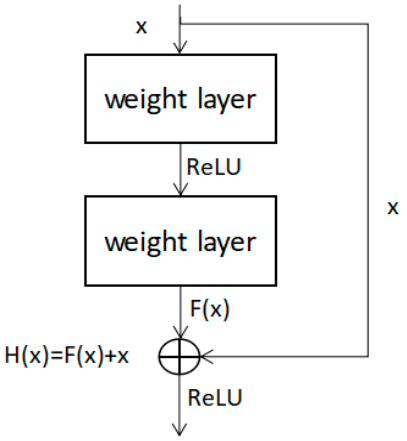

Figure 6. ResNet basic principle structure diagram.

Similarly, when it is coming to the *lth* layer, Equation (5) can be obtained.

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{5}$$

It can be seen that when the number of layers is deeper and deeper, the output value of the network is related to the output of the residual in the previous layer and the output of the Lth layer is the sum of the output values of the residuals of the previous layer. In back-propagation, calculation of the partial derivative of the loss function ε is as Equation (6).

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\sum_{i=l}^{L-1} F(x_i, W_i)\right) \tag{6}$$

It can be seen from the equation that the process of gradient derivation avoids the possibility that the value in the multiply state is 0, thus avoids the trouble caused by the disappearance of the gradient. After the network is improved in this way, even if the network is deeper, the results of network training will not be too deviated. This improved method avoids the problems of gradient disappearance and gradient explosion, which are caused by the increase in the number of network layers, which leads to maintaining a good network performance.

Taking the interactive video in the UT dataset [39] as an example, we used the Inception V3 pre-training model based on the ImageNet dataset. According to the requirements of the model, the image in the activity video was adjusted to a size of 299 × 299. The output of the last layer of the average pooling layer was used as the result of the preliminary feature information extraction of the Inception network. The results are stored as a series of feature files, each of which generates a feature file. Each segment of the video generates a feature file. During the experiment, each group of action videos was cut into 40 frames, so the output size of the network was 2048 × 40. During the experiment, for the ResNet pre-training network, the image was cropped to a uniform 224 × 224 size at the feature extraction stage due to the input data requirements of the pre-training model, Similarly, each video will generate a feature file, which is extracted based on the residual neural network with a feature size of 2048 × 40.

The proposed network is a multi-feature fusion convolutional neural network based on Inception and ResNet. At the feature extraction level, feature extraction is performed using the Inception V3 network and ResNet, respectively. Then, the two extracted features are fused at the fully connected layer. Then, training is continued. After multi-feature network convergence, the image features will be more abundant, which is conducive to improve the accuracy of human interaction video classification and recognition. According to the name of the video, this paper fused two feature files generated by each video file to form a feature file with a size of 4096 × 40 for later training and classification.

2.4. Whole-Individual Detection Based on Decision-Level Fusion

The decision-level fusion uses different features to obtain the classification results, and then the experimental results are merged. In the classification recognition phase, the whole video, individual video (left), and individual video (right) all produce a preliminary classification result based on probability. In order to make better use of the feature information of video images and improve the action recognition accuracy of interactive video, from the perspective of probability fusion, this paper fused the preliminary classification results at the decision level to obtain the final classification results. As shown in Equation (7), probabilistic weighting is used to fuse the three classification results of each set of action videos to obtain the final classification result:

$$R_{Final} = R_{Overall} \times P_{Overall} + R_{Left} \times P_{Left} + R_{Right} \times P_{Right} \quad (7)$$

Among them, R_{Final} is the final recognition result, R_{whole} is the whole video classification result of the double, R_{Left} represents the classification result after training using the individual video (left), and R_{Right} represents the classification result obtained by using the individual video (right) for model training. P_{whole}, P_{Left}, P_{Right} are the weighted probability of the whole video classification result of the two person, the weighted probability of the individual video (left) classification result, and the weighted probability of the individual video (right) classification result, respectively. The weighted probability value was obtained by comparing and analyzing the repeated experiments. A human body interactive video recognition block diagram based on decision level fusion is shown in Figure 7. As shown in Figure 7, after the video sequence is processed by a parallel multi-feature network, we can obtain the global and individual features. After the global feature and individual features are classified by Softmax, we can obtain a preliminary classification result based on probability. In the final classification stage, in order to make better use of the feature information of video images and improve the recognition accuracy of interactive video, this paper combined the preliminary classification results from the perspective of probability fusion to obtain the final classification results.

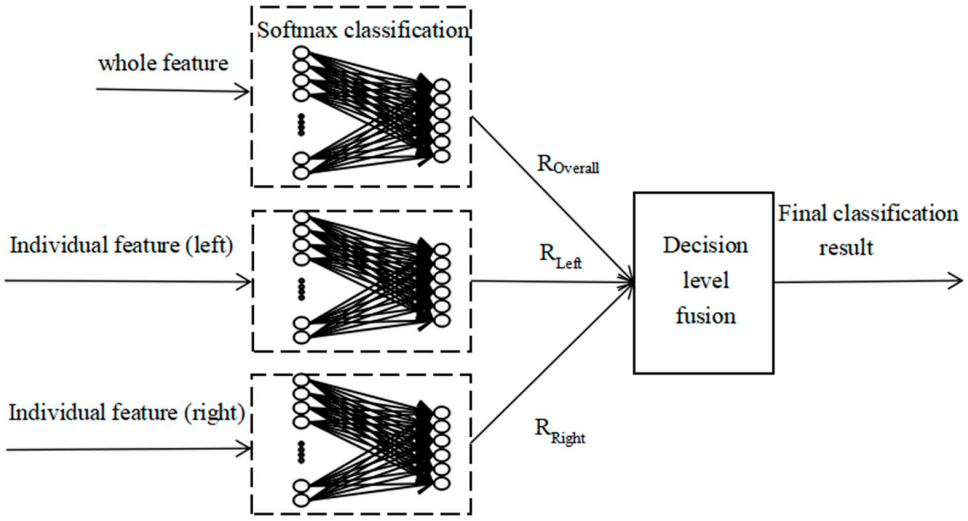

Figure 7. Human body interactive video recognition based on decision level fusion.

3. Results

3.1. Experimental Platform and Experimental Data

This paper conducted the experiments on a computer with a NVIDIA GTX 1080Ti graphics card. The experiment used Tensorflow for the deep learning network framework, Python for programming, MATLAB for other data analysis and preprocessing stages; the software used were Pycharm2018 and MATLAB2016.

This paper selected the UT interaction dataset. The UT dataset is a common human interaction video dataset. In order to ensure the effectiveness of the algorithm, the original data are generally divided into a training set and a verification set. This paper used K-fold cross validation (K-CV). This method divides the original data into K groups on average (K is generally greater than or equal to 2, the actual operation generally starts from 3, we only try to take 2 when the amount of raw data is small. Therefore, in this paper, we selected K = 3 for the experiment.) Each subset of data is used as a verification set, and the remaining K-1 subsets of data are used as training sets. In this way, K models are obtained. Finally, we used the average number of classification accuracy of the final validation set of these K models as the final result. This method is the most widely used as K-CV can effectively avoid the occurrence of over-fitting and under-fitting and the final result has high reliability. The UT-Interaction dataset contains videos of continuous executions of six classes of human–human interactions: shake-hands, point, hug, push, kick, and punch. Ground truth labels for these interactions are provided including time intervals and bounding boxes. There is a total of 20 video sequences whose lengths are around one minute. Each video contains at least one execution per interaction, providing us with eight executions of human activities per video on average. Several participants with more than 15 different clothing conditions appear in the videos. The videos are taken with the resolution of 720 × 480, 30 fps, and the height of a person in the video is about 200 pixels. We divided videos into two sets. Set 1 was composed of 10 video sequences taken on a parking lot. The videos of set 1 were taken with slightly different zoom rate, and their backgrounds were mostly static with little camera jitter. Set 2 (i.e., the other 10 sequences) were taken on a lawn on a windy day where the background is moving slightly (e.g., tree moves), and they contain more camera jitters. From sequences 1 to 4 and from 11 to 13, only two interacting people appear in the scene. From sequences 5 to 8 and from 14 to 17, both interacting people and pedestrians are present in the scene. In sets 9, 10, 18, 19, and 20, several pairs of interacting people execute the activities simultaneously. Each set has a different background, scale, and illumination. As shown in Figures 8 and 9, UT set 1 is an example of an action in the background of a parking lot, and UT set 2 was photographed on a windy lawn. For the entire video dataset, the video capture had different backgrounds, different resolutions, and different lighting conditions, which all bring challenges to the experiment of human interaction recognition.

In addition, in order to verify the applicability of this algorithm, we selected the interactive action video in UCF101 [40] for extended experiments. Since we are interested in human interactive videos, we selected a part of the interactive videos in the UCF101 dataset for training and testing. Our experiment on the UCF101 dataset was only to verify the effectiveness and versatility of the method. At present, the experiments of interactive action recognition are usually compared on the UT dataset. UCF101 is an action recognition dataset of realistic action videos collected from YouTube, having 101 action categories. This dataset is an extension of the UCF50 dataset, which has 50 action categories. With 13,320 videos from 101 action categories, UCF101 had the largest diversity in terms of actions and contained the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc., which made it the most challenging dataset to date. As most of the available action recognition datasets are not realistic and are staged by actors, UCF101 aims to encourage further research into action recognition by learning and exploring new realistic action categories. The videos in 101 action categories were grouped into 25 groups, where each group can consist of 4–7 videos of an action. The videos from the same group may share some common features such as similar background, similar viewpoint, etc. We randomly

divided the original data of the UCF101 interactive video into two groups: one as the training set and one as the verification set. We then used the training set to train the classifier, and then used the verification set to verify the model, and recorded the final classification accuracy as the result.

Figure 8. Example of UT set 1 dataset.

Figure 9. Example of the UT set 2 dataset.

More experimental parameters are as follows. We chose the Adam optimizer in the experiment, the learning rate = 10^{-5}, and decay = 10^{-6}. The loss function uses categorical cross entropy. During training, batch size = 32. During the training process, the image in the action video was adjusted to an image of size of 299 × 299. In the feature extraction stage, the image was cropped to a uniform size of 224 × 224. During the experiment, each group of action video clips was 40 frames, so the output size of the network was 2048 × 40.

3.2. Analysis of Results

In order to verify the improved effect of video acquisition based on Gaussian downsampling, we took the interactive video of UT dataset 1 as an example. We adopted equal interval downsampling video processing and Gaussian model based downsampling video acquisition to perform UT whole video preprocessing. In this paper, we selected the Inception network and ResNet for feature extraction and classification recognition processing, and the experimental results are shown in Table 1.

Table 1. Comparison of recognition accuracy of different sampling methods.

Sampling Method	Inception (%)	ResNet (%)	Multi-Feature Fusion (%)
Equal interval sampling	66.7	72.2	83.3
Downsampling based on Gaussian model	69.4	77.8	86.1
Equal interval sampling	66.7	72.2	83.3

It can be seen from Table 1 that the Gaussian model based downsampling method could improve the accuracy of human interaction recognition. The experimental process found that the improved fusion time-phased Gaussian model downsampling algorithm had an improved effect on the recognition of punching and pushing human interaction. However, the recognition and improvement of other actions were not obvious. This might be due to the limiting rules (some action video durations were shorter) of the video in the chosen dataset. Furthermore, there was no big difference in the choice of sampling methods.

In order to verify the human–interaction recognition algorithm based on the whole–individual detection proposed in this paper, a comparative experiment was carried out on the UT dataset. The UT interactive dataset generates individual video (left) and individual video (right) after the previous interactive video detection process. In this paper, UT individual video (left), UT individual video (right), and whole video were used for the experiments. In order to obtain reliable experimental classification and comparison, the experimental process was based on an Inception and ResNet parallel multi-feature fusion network algorithm. In the video downsampling process stage, we used a Gaussian model with improved fusion time phase features.

The experimental results of the UT dataset 1 were analyzed. Among them, the preliminary classification results of the individual actions obtained by the individual (left) video are shown in Figure 10, and the preliminary classification results obtained by the individual video (right) are shown in Figure 11. The preliminary classification results based on the whole video are shown in Figure 12. Finally, we combined the whole video and preliminary classification results of two groups of segmented video components for decision level fusion. The classification results of each action video after fusion are shown in Figure 13. In our experiments, the vertical axis of the confusion matrix is the actual actions, and the horizontal axis is the recognized actions.

The results of our experiments on the UT human interaction dataset are shown in Figures 14 and 15. It can be seen from Figure 14 that the accuracy of the training set and the validation set steadily increased, while Figure 15 is an image of the loss function. In terms of a specific sample, the loss function refers to the gap between the value predicted by the model and the true value. For a sample (x_i, y_i), y_i is the true value and $f(x_i)$ is our predicted value. Use the loss function $L(f(x_i), y_i)$ to represent the gap between the true value and the predicted value. At the same time, it can be seen from Figure 15 that the loss function of the training set and the validation set both steadily decreased. In the end, they tended to coincide.

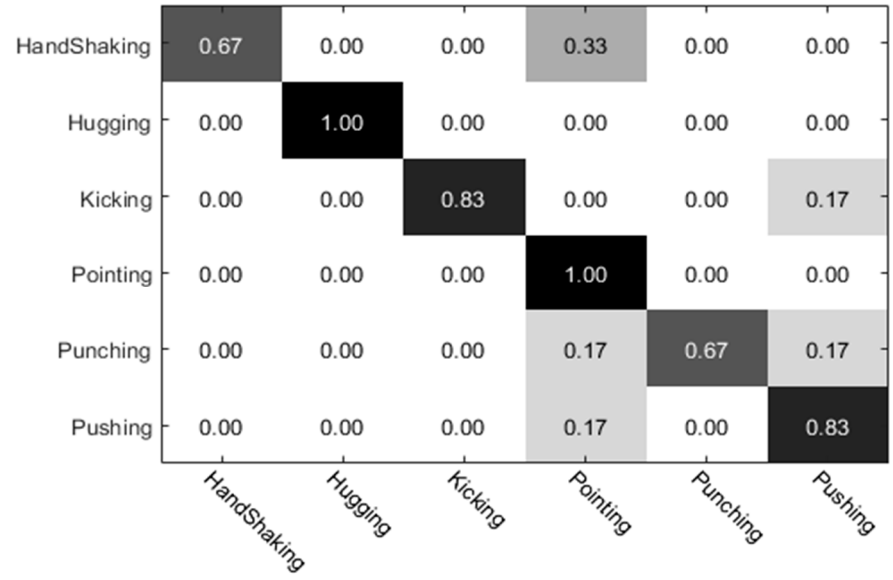

Figure 10. Individual video (left) classification results.

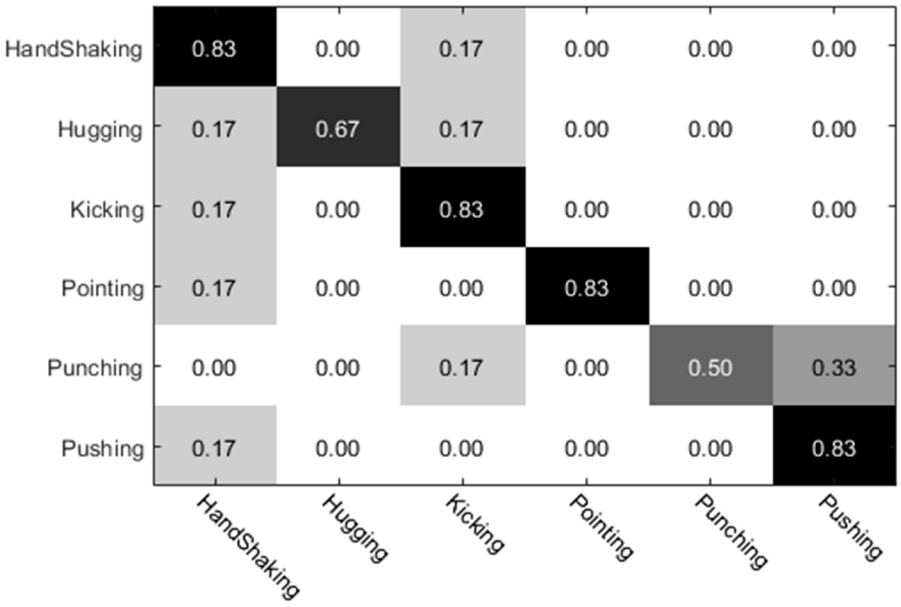

Figure 11. Individual video (right) classification results.

After the deciding level fusion, the whole accuracy of the video classification results is shown in Table 2. In the UT set 1 human interaction dataset, the individual video (left) obtained 83.3% recognition accuracy, the individual video (right) accuracy rate reached 75%, and the recognition accuracy based on UT whole video reached 86.1%. In contrast, the recognition result of the UT dataset was better than the video classification result after the individual split. The decline in recognition accuracy after individual detection may be due to the easy loss of feature information of interactive actions during video detection. In individual-split video, kicking, punching, and pushing actions always have a side action performer in an evasive state, thus the motion discrimination degree is small. Thus, the recognition accuracy of the individual split video is relatively low.

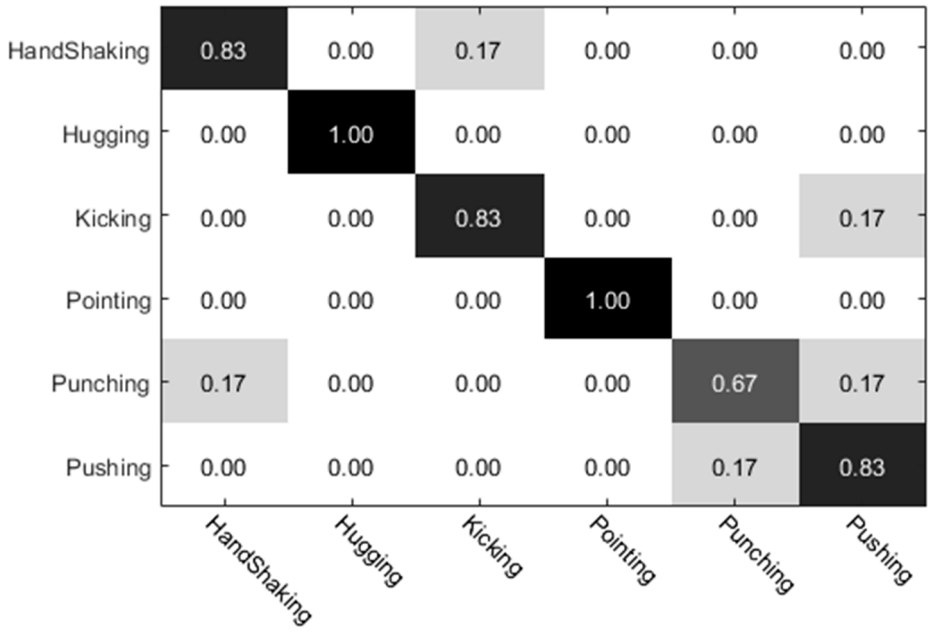

Figure 12. Whole video classification results.

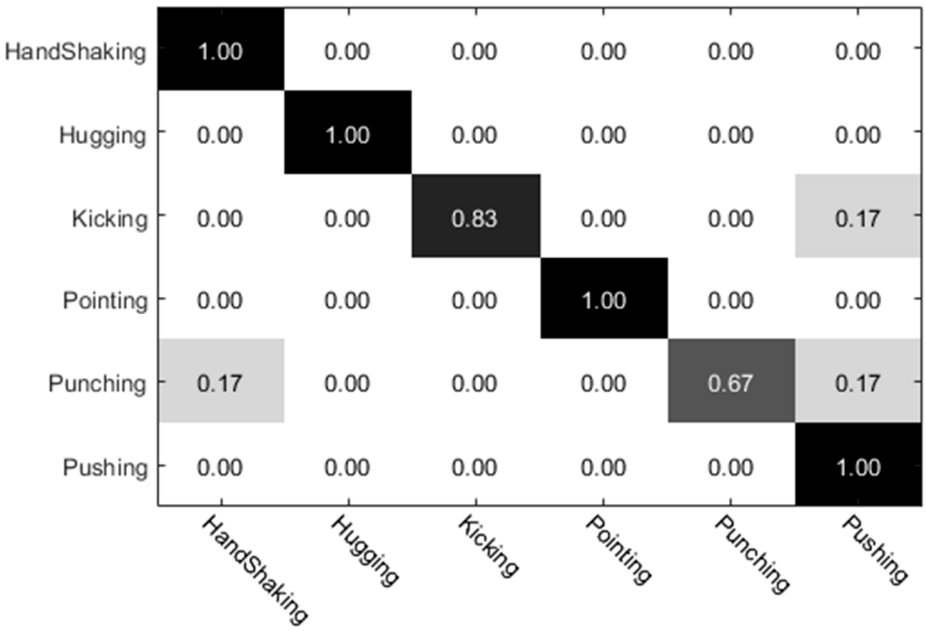

Figure 13. Classification results of human interaction based on whole–individual detection.

Table 2. Comparison of the individual video and whole video recognition accuracy.

Recognition Methods	UT Set 1 Recognition Accuracy (%)	UT Set 2 Recognition Accuracy (%)	UCF101 Interactive Recognition Accuracy (%)
Individual (left)	83.3	77.8	75.6
Individual (right)	75.0	72.2	76.8
whole	86.1	83.3	81.8
Fusion of this paper	91.7	86.1	85.4

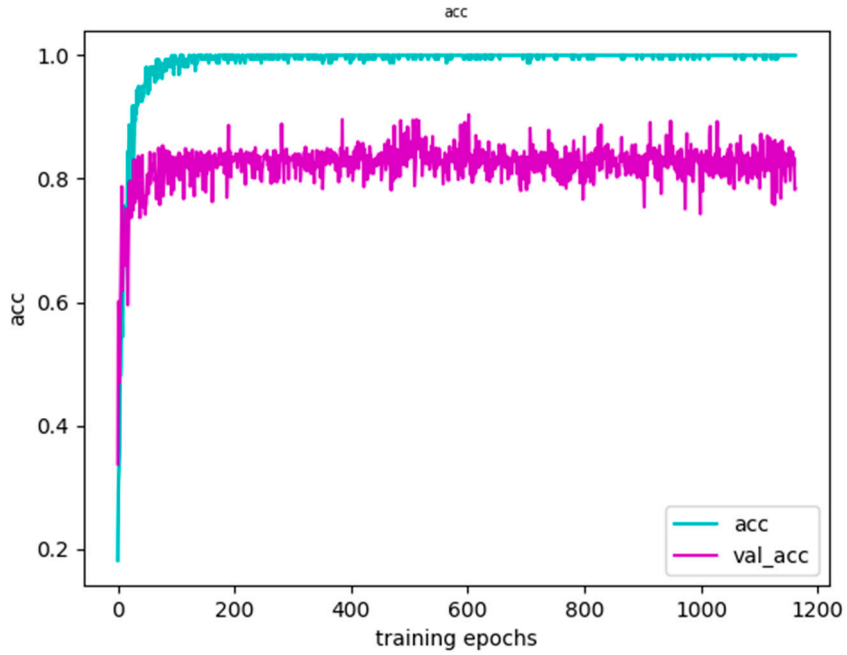

Figure 14. Accuracy of the training results.

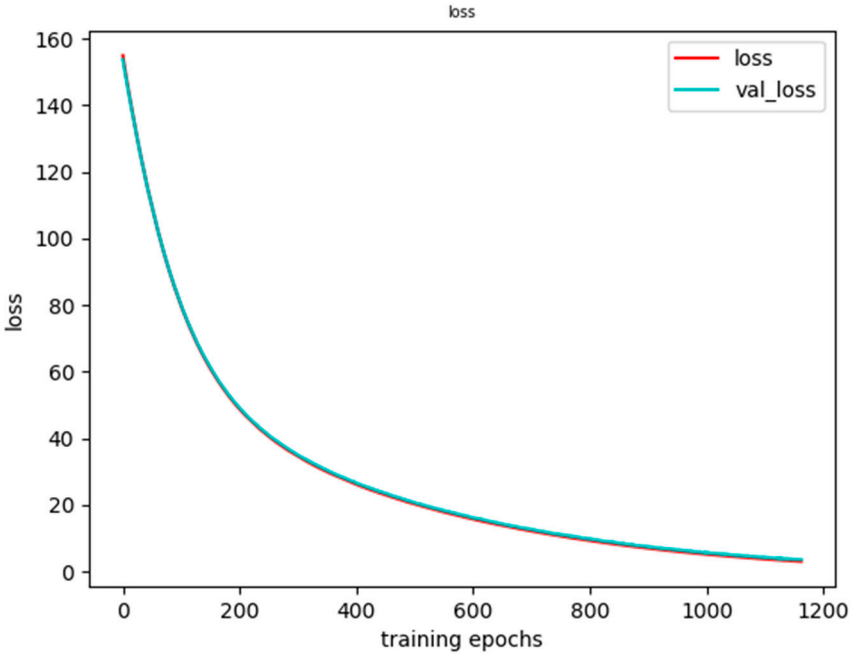

Figure 15. Loss function of the training results.

The comparison experiment results are shown in Table 2. For the selection of the optimal weights, we adopted the method of weight traversal. After the decision level fusion, the recognition accuracy of interactive video was significantly improved. From the analysis of the recognition accuracy of each action, it can be seen that the accuracy of the classification results of pointing and pushing was higher. This was not hard to predict since the characteristics of these two actions were more obvious. The classification result of the punching action needs to be improved because it is easily recognized as a push action and a handshake action. The kicking action was easily confused with the pushing action due to the low discrimination of degree of motion. The experimental results showed that when the detection video and the corresponding original whole video classification results were fused according to a certain probability, the classification accuracy of the handshake action and the push action was

significantly improved in the human interaction. In addition, in order to prove the credibility of this method, we also performed extended experiments on the UCF101 dataset. In the UCF101 human interaction dataset, the individual video (left) obtained a recognition accuracy of 75.6%, the individual video (right) obtained a recognition accuracy of 76.8%, and the UCF101 whole interactive video obtained a recognition accuracy of 81.8%. Thus, it can be seen that the method of individual detection and whole fusion proposed in this paper can improve the accuracy of human interaction recognition.

The experimental results of this paper were compared with the classification results of other experimental methods on the UT dataset in recent years, and the experimental results obtained are shown in Table 3. Huang et al. [30] used the HIS color space model to analyze the characteristics of the direction gradient histogram for different channels. Then, multi-channel fusion yielded 81.7% recognition accuracy. Mahmood et al. [41] proposed a new human interaction recognition (HIR) method that analyzes from local features, captures intensity changes, and distance from point to point. Time-based relationships identify key body points throughout the body contour and they used this method to extract the spatiotemporal characteristics of each different interaction. In their paper, the recognition in the UT set 1, two dataset experiments obtained an accuracy of 83.5% and 72.5%, respectively. Kong et al. [42] introduced an interactive phrase descriptor to represent the human interaction movement relationship and obtained a recognition rate of 88.33% in the UT dataset experiment. Shariat et al. [43] proposed a detection alignment model to improve the similarity measure between different action sequences, where they obtained a recognition rate of 91.57%. Guo et al. [44] proposed a new local descriptor based on the traditional descriptor LBP, and extended it to a space–time space. They obtained a recognition accuracy of 91.42% by using the neighborhood information of the three-dimensional cube. Using the method proposed in this paper, the experiment was carried out with the UT dataset, and we finally obtained a recognition rate of 91.70%. In order to prove the effectiveness of this method, we performed verification experiments on the UCF101 dataset and obtained a recognition rate of 85.43%. Therefore, it can be seen that the human-interaction recognition algorithm based on the whole–individual detection proposed in this paper can improve the accuracy of human interaction recognition rate.

Table 3. Comparison of the recognition accuracy of different identification methods.

Recognition Methods	UT Set 1 Recognition Accuracy (%)	UT Set 2 Recognition Accuracy (%)	UCF101 Interactive Recognition Accuracy (%)
HIS color space model [30]	81.70	–	–
Interactive phrases [42]	88.33	–	–
detection alignment model [43]	91.57	–	–
Novel 3D gradient LBP descriptor [44]	91.42	–	–
Method of this paper	91.70	86.10	85.43

4. Conclusions

In this paper, the human-interaction recognition algorithm based on whole–individual detection was proposed. The experimental verification and analysis work was carried out with the human interaction UT dataset. In the stage of feature information extraction of human interaction, we proposed human-interaction recognition in a parallel multi-feature fusion network. Compared with a single feature information extraction network, the fusion network improved the whole recognition accuracy of interactions. Regarding the complexity of interactive action features, we proposed an improved human body interaction recognition method based on the whole–individual detection. For the differences in action characteristics at different time periods, a Gaussian-based video downsampling method was proposed. This method makes the data acquisition more consistent with the characteristics of each action time. The results show that the whole–individual detection human interaction recognition method based on decision-level fusion proposed in this paper can improve the accuracy of classification

recognition. In this paper, the complexity of the algorithm was relatively high. Subsequent work will further improve the complexity of the algorithm to improve the efficiency of recognition.

Author Contributions: Conceptualization, H.Z., Q.Y., and C.Q.; Formal analysis, H.Z. and C.Q.; Funding acquisition, Y.Z.; Methodology, Q.Y. and C.Q.; Project administration, Y.Z.; Resources, C.Q. and Y.Z.; Software, H.Z. and C.Q.; Supervision, Q.Y. and Y.Z.; Validation, H.Z. and C.Q.; Writing—original draft, H.Z.; Writing—review & editing, H.Z. and Q.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Technology Project of Beijing Municipal Education Commission (No. SQKM201810009002), the National Natural Science Foundation of China (No .61371143), the National Natural Science Foundation of China (No. 61806008), and the Ministry of Education Science and Technology Development Center Project (No. 2018A03029).

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

References

1. Qi, J.; Jiang, G.; Li, G.; Sun, Y.; Tao, B. Intelligent Human-Computer Interaction Based on Surface EMG Gesture Recognition. *IEEE Access* **2019**, *7*, 61378–61387. [CrossRef]
2. Minhaz, U.A.; Yeong, H.K.; Jin, W.K.; Md, R.B.; Phill, K.R. Two person Interaction Recognition Based on Effective Hybrid Learning. *KSII Trans. Int. Inf. Syst.* **2019**, *13*, 751–770.
3. Chinimilli, P.T.; Redkar, S.; Sugar, T. A Two-Dimensional Feature Space-Based Approach for Human Locomotion Recognition. *IEEE Sens. J.* **2019**, *19*, 4271–4282. [CrossRef]
4. Phyo, C.N.; Zin, T.T.; Tin, P. Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints. *IEEE Trans. Consum. Electron.* **2019**, *65*, 243–252. [CrossRef]
5. Joao, C.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
6. Qi, H.; Fang, K.; Wu, X.; Xu, L.; Lang, Q. Human activity recognition method based on molecular attributes. *Int. J. Distrib. Sens. Netw.* **2019**. [CrossRef]
7. Sanzari, M.; Ntouskos, V.; Pirri, F. Discovery and recognition of motion primitives in human activities. *PLOS ONE* **2019**, *14*, e0214499. [CrossRef]
8. An, F. Human Action Recognition Algorithm Based on Adaptive Initialization of Deep Learning Model Parameters and Support Vector Machine. *IEEE Access* **2018**, *6*, 59405–59421. [CrossRef]
9. McColl, D.; Jiang, C.; Nejat, G. Classifying a Person's Degree of Accessibility from Natural Body Language During Social Human–Robot Interactions. *IEEE Trans. Cybern.* **2017**, *47*, 524–538. [CrossRef] [PubMed]
10. Wang, Z.; Cao, J.; Liu, J.; Zhao, Z. Design of human-computer interaction control system based on hand-gesture recognition. In Proceedings of the 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Hefei, China, 19–21 May 2017; pp. 143–147.
11. Lakomkin, E.; Zamani, M.A.; Weber, C.; Magg, S.; Wermter, S. On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 854–860. [CrossRef]
12. Böck, R. Recognition of Human Movement Patterns during a Human-Agent Interaction. In Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI'18), New York, NY, USA, 16 October 2018; pp. 33–37. [CrossRef]
13. Lou, X.; Yu, Z.; Wang, Z.; Zhang, K.; Guo, B. Gesture-Radar: Enabling Natural Human-Computer Interactions with Radar-Based Adaptive and Robust Arm Gesture Recognition. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 9 October 2018; pp. 4291–4297. [CrossRef]
14. Faria, D.R.; Vieira, M.; Faria, F.C.C.; Premebida, C. Affective facial expressions recognition for human-robot interaction. In Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal, 28 August 2017; pp. 805–810. [CrossRef]
15. Käse, N.; Babaee, M.; Rigoll, G. Multi-view human activity recognition using motion frequency. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 18–20 September 2017; pp. 3963–3967. [CrossRef]

16. Jaouedi, N.; Boujnah, N.; Htiwich, O.; Bouhlel, M.S. Human action recognition to human behavior analysis. In Proceedings of the 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Hammamet, Tunisia, 18–20 December 2016; pp. 263–266. [CrossRef]
17. Silambarasi, R.; Sahoo, S.P.; Ari, S. 3D spatial-temporal view based motion tracing in human action recognition. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), Wuhan, China, 17–19 March 2017; pp. 1833–1837. [CrossRef]
18. Tozadore, D.; Ranieri, C.; Nardari, G.; Guizilini, V.; Romero, R. Effects of Emotion Grouping for Recognition in Human-Robot Interactions. In Proceedings of the 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, 22–25 October 2018; pp. 438–443. [CrossRef]
19. Liu, B.; Cai, H.; Ji, X.; Liu, H. Human-human interaction recognition based on spatial and motion trend feature. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 4547–4551. [CrossRef]
20. Wang, H.; Wang, L. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3633–3642. [CrossRef]
21. Zhao, Y.; Xiong, Y.; Lin, D. Trajectory convolution for action recognition. *Adv. Neural Inf. Processing Syst.* **2018**, *2018*, 2205–2216.
22. Chiang, T.; Fan, C. 3D Depth Information Based 2D Low-Complexity Hand Posture and Gesture Recognition Design for Human Computer Interactions. In Proceedings of the 2018 3rd International Conference on Computer and Communication Systems (ICCCS), Nagoya, Japan, 27–30 April 2018; pp. 233–238. [CrossRef]
23. Du, T.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
24. Vox, J.P.; Wallhoff, F. Preprocessing and Normalization of 3D-Skeleton-Data for Human Motion Recognition. In Proceedings of the 2018 IEEE Life Sciences Conference (LSC), Montreal, QC, Canada, 28–30 October 2018; pp. 279–282. [CrossRef]
25. Phyo, C.N.; Zin, T.T.; Tin, P. Skeleton motion history based human action recognition using deep learning. In Proceedings of the 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), Nagoya, Japan, 24–27 October 2017; pp. 1–2. [CrossRef]
26. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. Multi-fiber networks for video recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
27. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
28. Li, N.J.; Cheng, X.; Guo, H.Y.; Wu, Z.Y. Recognizing human interactions by genetic algorithm-based random forest spatio-temporal correlation. *Pattern Anal. Appl.* **2016**, *19*, 267–282. [CrossRef]
29. Huang, F.F.; Cao, J.T.; Ji, X.F. Two-person interactive motion recognition algorithm based on multi-channel information fusion. *Comput. Technol. Dev.* **2016**, *26*, 58–62.
30. Guo, P.; Miao, Z.; Zhang, X.; Shen, Y.; Wang, S. Coupled Observation Decomposed Hidden Markov Model for Multiperson Activity Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1306–1320. [CrossRef]
31. Ji, X.F.; Wang, C.H.; Wang, Y.Y. A two-dimensional interactive motion recognition method based on hierarchical structure. *J. Intell. Syst.* **2015**, *10*, 893–900.
32. Vahdat, A.; Gao, B.; Ranjbar, M.; Mori, G. A discriminative key pose sequence model for recognizing human interactions. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1729–1736. [CrossRef]
33. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2015. [CrossRef]
34. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
35. Kalman, R. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7 June 2015. [CrossRef]

37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
38. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
39. Ryoo, M.S.; Chen, C.C.; Aggarwal, J.K.; Roy-Chowdhury, A. An Overview of Contest on Semantic Description of Human Activities (SDHA) 2010. In *Recognizing Patterns in Signals, Speech, Images and Videos. ICPR 2010. Lecture Notes in Computer Science*; Ünay, D., Çataltepe, Z., Aksoy, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6388. Available online: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html (accessed on 26 August 2010).
40. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild. CRCV-TR-12-01. November 2012. Available online: http://crcv.ucf.edu/data/ (accessed on 1 November 2012).
41. Mahmood, M.; Jalal, A.; Sidduqi, M.A. Robust Spatio-Temporal Features for Human Interaction Recognition Via Artificial Neural Network. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018; pp. 218–223. [CrossRef]
42. Kong, Y.; Jia, Y.; Fu, Y. Learning Human Interaction by Interactive Phrases. In *Computer Vision—ECCV 2012. ECCV 2012. Lecture Notes in Computer Science*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7572.
43. Shariat, S.; Pavlovic, V. A New Adaptive Segmental Matching Measure for Human Activity Recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 3–6 December 2013; pp. 3583–3590. [CrossRef]
44. Guo, Z.; Wang, X.; Wang, B.; Xie, Z. A Novel 3D Gradient LBP Descriptor for Action Recognition. *Trans. Inf. Syst.* **2017**, *100*, 1388–1392. [CrossRef]