3D Behavior Recognition Based on Multi-Modal Deep Space-Time Learning

Chong Zhao ¹, Minglin Chen ¹, Jinhao Zhao ¹, Qicong Wang ¹ and Yehu Shen ²*

¹ Department of computer science, Xiamen University, Xiamen 361005, China; zhc@xmu.edu.cn (C.Z.); chenminglin@stu.xmu.edu.cn (M.C.); zhaojinhao0592@163.com (J.Z.); qcwang@xmu.edu.cn (Q.W.)
² College of mechanical engineering, Suzhou University of Science and Technology, Suzhou 215009, China
* Correspondence: yehushen@mail.usts.edu.cn; Tel.: +86-512-6831-0016

Received: 23 December 2018; Accepted: 15 February 2019; Published: 19 February 2019

Abstract: This paper proposes a dual-stream 3D space-time convolutional neural network action recognition framework. The original depth map sequence data is set as the input in order to study the global space-time characteristics of each action category. The high correlation within the human action itself is considered in the time domain, and then the deep motion map sequence is introduced as the input to another stream of the 3D space-time convolutional network. Furthermore, the corresponding 3D skeleton sequence data is set as the third input of the whole recognition framework. Although the skeleton sequence data has the advantage of including 3D information, it is also confronted with the problems of the existence of rate change, temporal mismatch and noise. Thus, specially designed space-time features are applied to cope with these problems. The proposed methods allow the whole recognition system to fully exploit and utilize the discriminatory space-time features from different perspectives, and ultimately improve the classification accuracy of the system. Experimental results on different public 3D data sets illustrate the effectiveness of the proposed method.

Keywords: space-time characteristics; feature fusion; 3D skeleton

1. Introduction

As an important branch of computational vision, behavior identification has a wide range of applications, such as intelligent surveillance, medical care, human-computer interaction, virtual reality, etc. [1,2]. With the advent of RGB-D (RGB-Depth) sensor technology, behavior identification is able to capture both the RGB image of the scene and its corresponding depth map simultaneously in real time. Depth maps can provide 3D geometric cues that are less sensitive to illumination variations compared with traditional RGB images [3,4]. Furthermore, the depth sensor can provide real-time estimations of the 3D joint positions of the human skeleton, and can further eliminate the effects of cluttered background and brightness changes. Shotton et al. [5] proposed a very powerful human motion capture technique that could estimate the 3D joint position of a human skeleton from a single depth map. Since then, 3D human behavior recognition with RGB-D multimodal sequence data has drawn great attention in recent years.

The problem of 3D behavior recognition is defined as follows. A set of depth images $D_I = \{D_1, D_2, ..., D_n\}$ are captured by RGB-D camera (such as Microsoft Kinect) and a set of predefined behavior categories $L = \{l_1, l_2, ..., l_m\}$. Our aim is to find an optimal mapping $y = f(D_I)$ from $D_I$ to $L$ so that we can estimate the behavior category from the depth sequence.

Current 3D behavior recognition could be roughly divided into 2 categories: feature extraction with handcrafted methods and deep learning based methods. Wang et al. [6] designed a 3D local occupancy pattern feature to describe the local depth appearance at the joint locations for interacting with the target. The method calculates the number of points falling into space-time bin when the
space around the joint is occupied by the target. In follow-up work, Wang et al. [7] used the 3D point cloud around special joint points to calculate the LOP features and identify different types of interactions. They also applied Fourier timing pyramid to represent the timing structure. Based on the above two types of features, the Actionlet Ensemble model was proposed in reference [7]. It incorporates the features from a subset of joints. Althloothi et al. [8] proposed two feature sets: the shape representation features from the skeleton data and the features of the kinematic structure in order to fuse depth-based features and skeleton-based features. Shape features are used to describe 3D joint contour structures while kinematic features are used to describe human motion. These two feature sets are used for behavior recognition through multicore learning techniques to fuse features at the kernel level. Similarly to reference [8], Chaaraoui et al. [9] proposed a fusion approach which combined skeleton-based and contour-based features. Skeleton information is crucial for 3D behavior recognition. Most of the existing skeleton-based behavior identification methods use Hidden Markov Models (HMMs) [10–12] and Time Series Pyramids (TPs) [13–15] to accurately construct the temporal dynamics of skeleton joints. It is difficult for HMMs to obtain a consistent sequence of timing and the corresponding distribution. TPs methods usually adopt a variable time window and can obtain reliable and distinguishable skeleton space-time characteristics using time-domain context information. In reference [16], the authors obtained various STIP (Spatiotemporal Interest Point) features by combining detectors with different feature points and comparing them with the traditional STIP method. They showed that the skeleton joint information could greatly enhance the performance of the depth-based behavior recognition method by fusing time-space features with skeleton joint points [16]. Recently, Chen et al. presented a local spatial-temporal descriptor based on local binary patterns histogram features for action recognition from depth video sequences [17]. Generally speaking, more reliable 3D behavior recognition with RGB-D sensor data could be achieved by carefully designing and integrating the distinguishable space-time features from depth map sequences and skeleton sequence data. However, most of the above-mentioned approaches only use handcrafted designing methods. Although the methods with handcrafted features do not require huge amount of training data which is the case for deep learning based methods. It is well known that handcrafted features are general methods and are not optimal for specific tasks such as 3D behavior recognition. Furthermore, handcrafted features could not fully exploit the intrinsic space-time information contained in the RGB-D sequence so that their performances are not satisfactory.

Current deep neural network techniques demonstrate better performances in automatic space-time feature learning [18]. Deep convolutional framework has become an effective method to extract high-level space-time features from the original RGB-D depth data for 3D behavior identification [19]. On the other hand, 3D behavior recognition using skeletal sequence data generated by RGB-D is generally regarded as a time series problem [20,21], which is to say the body can be represented by extracting body posture and kinetic characteristics over time. Current methods mainly focus on using only single data sources as input, such as either using depth sequences or using 3D skeleton data so that their performances are still limited. In fact, different data sources are usually complementary to each other. It is beneficial to combine different data sources in order to further boost performances. Therefore, a 3D space-time convolutional network is proposed based on the previous work to automatically learn two kinds of 3D modal data from RGB-D depth map sequences (spatial-temporal convolutional features of depth map cubes and depth motion maps) for 3D behavior recognition. We also combine the space-time characteristics of the skeleton joints based on temporal pyramids to further improve the accuracy of 3D behavior recognition.

As shown in Figure 1, our proposed human behavior identification framework mainly consists of three parts: feature learning based on 3D spatial-temporal convolutional neural networks, human skeleton representation and the integration of classification results. During the feature learning phase, 3D space-time convolutional neural network is used to extract the space-time features from the original depth video sequence and depth motion frame sequence. The output of the last layer is extracted to be the feature from a 3D space-time convolutional neural network. In the human skeleton representation
phase, the relative three-dimensional geometric relationships between the body parts in the human skeleton are calculated, and then the human motion can be modeled as a curve in the manifold $SE(3) \times \ldots \times SE(3)$. Finally, the high-level features extracted from the 3D space-time convolutional neural network and the SVM classification results which are represented by the human skeleton manifold are fused together to achieve the behavior recognition. For the fusion process, we multiply the probabilities of the corresponding actions from three SVM classifiers with different weights, and the one with the highest value is the identified action class.

![Figure 1. Flowchart of the entire fusion framework. (3DSTCNN in blue blocks stands for 3D spatial-temporal convolutional neural networks).](image)

The main contributions of this paper are as follows.

1. In order to fully exploit and utilize the discriminatory space-time features from different perspectives, a fusion based action recognition framework which automatically learns advanced features directly from the original depth sequence and depth motion maps, and seamlessly integrates them with the human skeleton manifold representation is proposed in this paper.

2. The proposed method was evaluated on three popular public benchmark datasets, namely, MSR-Action3D, UTD-MHAD and UTKinect-Action3D datasets, and achieved state-of-the-art recognition results.

2. Related Work

2.1. Behavior Recognition based on Space-Time Convolution Features

In recent years, with the successful application of deep learning in image recognition, deep neural network based methods have gradually emerged in the area of human motion recognition. In the early stage of the study of these methods, Ref. [22] proposed an image feature expression that gradually increases in complexity and invariability as the levels increase based on the HMAX model. It was applied to the real scenes for target recognition. Its performance is obviously superior to traditional feature transformations (such as SIFT features etc.). In reference [23], a video recognition framework based on HMAX (Hierarchical Model and X) architecture was proposed with predefined space-time filtering in the first layer of the model. Based on reference [23], Ref. [24] formed a spatial-temporal identification flow with the help of a spatial HMAX model, which used the flow as an artificially tailored and relatively shallow (3-layer) HMAX model. In reference [25], unconstrained learning of space-time features was performed using a convolution-restricted Boltzmann machine and independent subspace analysis. Then it was embedded into a discriminative model to realize motion recognition. However, it still used low-level features. Most of the above feature learning methods still follow the feature learning framework designed for the single image. They ignore the temporal structure in the context of
the video frame. Video analysis studies based on convolutional neural networks in recent years show that better performance can be achieved by learning video temporal features. In motion recognition, the use of human-customized features for video analysis is gradually replaced by deep convolutional feature learning strategies. In reference [26], the convolutional neural network used for 2D image recognition was extended to 3D space by adding the time dimension, and multiple convolutional layers and down-sampling layers were used to learn the depth characteristics from the video data. The following linear classification model effectively identified the actions. Ref. [27] proposed a two-stream convolutional framework that used image and optical flow data as input. The convolutional neural network is used to extract the feature of the dense optical flow field in the video, and the single-frame background image is used to assist the action classification. Experimental results proved the improved performances. Ref. [28] proposed to divide the video into two kinds of data streams, including the detail part of the center video stream and the entire video of the under-sampling global data stream. These 2 data streams were processed by two convolutional neural networks with the same structure in parallel for video classification. Ref. [29] learned space-time filtering in a deep network and applied it to a variety of human motion understanding tasks. Wang et al. [30] extracted the appearance and motion features from the video sequence separately, combined them through a cascaded layer, and then proposed a time domain pyramid pooling strategy to implement motion recognition for variable-length videos. It can be concluded that this pooling strategy is an important way to enhance the expression of video motions through convolutional neural networks to implement joint mining of multiple spatial and temporal information from video sequences. The video data can be described effectively by learning from different angles such as temporal and spatial features.

2.2. Behavior Recognition based on Skeleton Space-Time Expression

Given a corresponding 3D joint position, a behavior can be represented by a time-domain sequence of a human skeleton that is close to it. Existing joint-based behavioral recognition methods use different features to model the motion of a joint or a group of joints. Various features are extracted from the skeleton data in order to obtain the relevance of body joints in 3D space. Yang et al. [31] proposed a new descriptor called EigenJoints which contained the pose, motion and offset features. The differences between the current frame and neighboring frames were used to encode the temporal and spatial information. Wang et al. [32] estimated human joint positions from the video and then divided the estimated joint positions into five different sections. Each action can be expressed by calculating the changes in the spatial and temporal parameters of these body parts. The author applied the bag-of-words method for classification. In reference [33], the behavior was modeled as a space-time structure that used a set of 13 joint locations in a 4D space (3D space plus time) to describe a human behavior. In reference [10], Lv et al. designed seven feature vectors, such as joint position, bone position, and joint angle etc. Then a Hidden Markov Model was used to learn the movement of each action for each feature. Chaudhry et al. [34] proposed a space-time hierarchical skeleton configuration in which each pose represented the motion of a set of joints in a special time domain. In [35], the joint-space covariance matrix in a fixed-length time window was used to obtain the space-time relationships between joints. Zanfir et al. [36] proposed a descriptor called motion gesture which was formed by the position, velocity, and acceleration of a skeleton joint in a short time window around the current frame. They considered both the spatial attitude information and the time difference of the human joint. The work in reference [37] proposed a more complex skeletal representation that described the relative geometry between two rigid limbs as the combination of a rigid body rotation and translation. Thus, the skeletal sequence could be expressed as a curve in the Lie group, which in turn used a Fourier temporal pyramid [38] to represent skeleton space-time features. In reference [39], each behavior was represented by the space-time motion trajectory of the joint, and the trajectory was a curve in the Riemannian manifold of an open-curve shape space. In general, skeleton space-time expression-based methods usually need intricate feature designs to classify actions without joint selection, and all the joints are all treated the same way during behavior identification.
3. Multi-Modal 3D Space-Time Convolutional Framework

3.1. Space-Time Feature Learning of Depth Sequences

The proposed 3D deep convolutional neural network is designed based on the previous work [40]. It contains two 3D convolution layers, each of which is followed by a 3D max pooling layer. The input to the network is a cube obtained by simply preprocessing the original depth sequence and normalizing the input to the same size as the training data set. Since the length of the action video is not the same in different data sets, it is difficult to normalize all the videos into the same size while still maintaining sufficient information in the video. Therefore, the sizes of cubes that are fed into the network from different data sets are different. Let’s take the MSR-Action3D dataset as an example. Firstly, after a simple pre-processing of the original depth sequence, a normalized depth cube (motion video) with dimension $32 \times 32 \times 38$ (height $\times$ width $\times$ time) is fed into a 3D deep convolutional neural network. The input cube is processed by the first 3D convolutional layer (C1) and the 3D maximum pooling layer (M1), followed by the second 3D convolutional layer (C2) and the 3D maximum pooling layer (M2). Finally, a fully connected layer and a soft-max layer are applied for classification. The kernel sizes of the convolutional layers C1 and C2 are $5 \times 5 \times 7$ and $5 \times 5 \times 5$, respectively. In the subsequent 3D maximum pooling layer of the 3D convolutional layer, $2 \times 2 \times 2$ downsampling is used in the M1 and M2 layers in order to reduce the space and time dimensions. The numbers of feature maps in C1 and C2 are 16 and 32, respectively. For the UTD-MHAD data set, the cube size of the input layer is $32 \times 32 \times 28$. Both of the convolution kernels in C1 and C2 are $5 \times 5 \times 5$. Unless otherwise specified, all other parameters in this paper are the same as those used in the MSR-Action3D dataset.

3.2. Space-Time Feature Learning of Depth Motion Map

Each depth frame in depth video sequence is projected onto 3 orthogonal Cartesian coordinate planes. In each projection plane, the entire depth video sequence accumulates the differences between 2 consecutive projections, and finally forms depth motion maps (DMM). Depth maps can be used to extract 3D structure and shape information. Yang et al. [41] proposed to project the depth frame onto three orthogonal Cartesian coordinate planes in order to describe the process of the motion. Because of its simplicity, we adopt similar ideas as reference [41] in this paper, but we also make some changes in the process of obtaining DMM. Specifically, each 3D depth frame can generate three 2D projection maps corresponding to the front, side, and top viewing angles, which are represented by $\text{map}_v$, where $v \in \{f, s, t\}$. For a point $(x, y, z)$ in the depth frame, where $z$ represents the depth value, the pixel values in three projection maps are indicated by $x, y$ and $z$, respectively. Different from [41], for each projection graph, the motion energy is obtained by computing the absolute difference of two consecutive frames without binarization. For a depth video sequence with $N$ frames, $\text{DMM}_v$ is obtained by stacking the motion energy of the entire depth video sequence as follows:

$$\text{DMM}_v = \sum_{i=a}^{b} |\text{map}^i_v - \text{map}^{i-1}_v|$$

where $i$ denotes the index of the video frame; $\text{map}^i_v$ denotes the projection of the $i$-th frame at the projection direction of $v$; $a \in \{2, \ldots, N\}$ and $b \in \{2, \ldots, N\}$ denote the indices of the starting frame and the ending frame. It is worth noting that not all frames in a depth video sequence are used to generate a DMM. After obtaining the DMM of the depth video sequence, the bounding-box is set to extract the non-zero area of the DMM and served as the foreground.

The foreground extracted from the DMM is represented as $\text{DMM}_v$. The DMM obtained from the three projection directions can effectively describe the movement in different ways. This is why DMM is used for behavioral identification in this paper. Because the size of the $\text{DMM}_v$ of different video sequences may be different, we normalize all the size of $\text{DMM}_v$ to $64 \times 64$ so that it is able to feed the extracted DMM into the 3D deep convolutional neural network for training. Since in our method the
feature is extracted from each pixel in the DMM image, the pixel values of DMM image are normalized into the range of 0~1 to avoid too large pixel values dominating the entire feature set.

The process of extracting the high level features from DMM is similar to the one mentioned in the previous section. The network also contains two 3D convolution layers, with each 3D volume being followed by a 3D maximum pooling layer. The size of the input cube for the network is $64 \times 64 \times 3$. Here we will transform different data sets which are fed into the cube of the network to the same size, namely $64 \times 64 \times 3$. Taking the MSR-Action3D data set as an example, the sizes of convolution kernels in C1 and C2 layers are $6 \times 6 \times 2$ and $5 \times 5 \times 1$ respectively, and the stride is 1. A 3D max pooling layer follows the 3D convolutional layer. The downsampling layer in the M1 and M2 layers has the size of $2 \times 2 \times 1$. The numbers of features in C1 and C2 layers are 16 and 32 respectively.

3.3. Implementation of the Hybrid Model

In order to process the original skeleton sequence data better, a series of pre-processing is performed on the original skeleton sequence so that it is more convenient for us to use. First of all, in order to easily manipulate the skeleton data, the human skeleton model is established. An index is assigned to each joint point, and the total numbers of joint points and body parts are given beforehand. The body part is described by the index pair of the two joint points. Take the UTKinect-3DAction dataset as an example, the index of the left shoulder joint is No. 1 and the index of the left elbow joint is No. 8. As such, the “left upper arm” of the body part can be represented by an index pair (1, 8). Other body parts can also be represented in a similar manner. With the above-mentioned pre-processing step, an $n \times 2$ matrix of all body parts is obtained, where $n$ represents the number of body parts. Each row in the matrix is an index pair of two body joints, that is, an index pair of a body part. Furthermore, in order to facilitate the following manifold representation, any two body parts are needed to be described by an index pair. Thus, an $N \times 4$ relative body part index matrix is formed, where $N = n \times (n - 1)$. Each row of the matrix is an index pair which describes a pair of body parts. Finally, in order to facilitate the comparison of the skeleton, we set the skeleton of a specific frame as a reference and normalize the length of the body parts in other skeleton data of the activity so that the length of the same body part in each frame is the same. With the above-mentioned procedure, the skeleton is invariant to the scale. The original skeleton sequence is further rotated around the hip center joint which is set as the origin of the coordinates so that the projection of the line which connects the left hip joint point and the right hip joint point in the xy plane is parallel to the x axis. The transformed skeleton is visually invariant and is used in the experiments. Comparisons before and after skeleton preprocessing are shown in Figure 2.

![Figure 2. Comparisons of the skeleton before and after preprocessing: (a) Before preprocessing; (b) After preprocessing.](image)

In the feature learning phase, the 3D space-time convolution model is used for the input sequence data of different modalities. Negative log-likelihood criteria are used in the optimization phase
which requires the output of the trainable model to be a suitable normalized logarithmic probability. Classification can be done by using the soft-max function. In this paper, the stochastic gradient descent algorithm is used to train the neural network.

We train an SVM classifier with the features from the depth image sequence, the space-time convolution feature of the depth motion map and the Fourier feature of the 3D human skeleton. During the test phase, the corresponding features are extracted from each action in the video, and fed into the corresponding trained model. Thus, the classification probability of each action category is estimated. Finally, the probabilities of the corresponding actions obtained with three models are fused according to linear combination. In this paper, the weights of linear combination are set to be 3, 3 and 5 respectively according to cross validation.

4. Experimental Results
4.1. Datasets

In this paper, experiments are conducted on three datasets: the UTD-MHAD dataset [42], the UTKinect-Action3D dataset [43] and the MSR-Action3D dataset [44]. All three datasets were captured by a fixed Microsoft Kinect sensor. All of the datasets provide the 3D coordinates of the corresponding joints. The results of our proposed method are compared against the current state-of-the-art on all 3 datasets.

The UTD-MHAD dataset was collected in an indoor environment. The dataset contains 27 actions which were performed by 8 actors (4 males and 4 females). Each actor performed 4 times for each action. This dataset includes 3D coordinates of 20 human joints. The dataset contains 861 image sequences.

The UTKinect-Action3D dataset consists of 10 actions, each of which was performed by 10 different actors. This dataset contains 3D coordinates of 20 human body joints. These 10 actions are “walk”, “sit down”, “stand up”, “pick up”, “lift”, “throw”, “push”, “pull”, “wield arm” and “clap”. Each actor executed twice for each action. The dataset contains 199 valid action sequences. For the sake of processing convenience, a total of 200 motion sequences were used in this paper. The missing “lift” motion of the tenth actor was interpolated with the 1242th to the 1300th frames. The main challenge of UTKinect-Action3D dataset is the perspective and the large intra-class variations.

The MSR-Action3D dataset contains 20 actions which were performed by 10 different actors. Each actor performed 2 or 3 times for each action. This dataset contains 3D coordinates of 20 human joints. These 20 actions are “high arm swing”, “level arm swing”, “hammer strike”, “grip”, “forward punch”, “high pitch”, “drawing X”, “painted hook”. “circle”, “high clap”, “waving hands”, “flank”, “bend”, “kick”, “side kick”, “jog”, “tennis swing”, “tennis serve” “golf swing” and “throw after picking”. There are 567 sequences in this dataset. According to the test environment of the current method [43], we divided these 20 actions into three action subsets AS1, AS2 and AS3 (as shown in Table 1). Each subset contains eight actions. The actions in AS1 and AS2 are a group of actions with similar motions, and the actions in AS3 are a complex set of actions. The main challenge in the MSR-Action3D dataset is that some of the actions are similar, especially in AS1 and AS2.

| AS1          | AS2          | AS3          |
|--------------|--------------|--------------|
| horizontal arm wave | high arm wave | high throw   |
| hammer       | hand catch   | forward kick |
| forward punch| draw X       | side kick    |
| high throw   | draw tick    | jogging      |
| hand clap    | draw circle  | tennis swing |
| bend         | two hand wave| tennis serve |
| tennis serve | forward kick | golf swing   |
| pickup and throw | side boxing | pickup and throw |

Table 1. Actions in actions subset AS1, AS2 and AS3.
In order to reduce the sensitivity of our proposed method against different testing sequences, we propose a preprocessing step for each depth image sequence. It mainly includes the foreground extraction, body-centered bounding box detection, normalization of spatial and temporal dimensions and normalization of depth values. After foreground extraction, only the depth values of the foreground pixels are preserved. Other pixels are set to zero. In the process of clipping the bounding box, the maximum values of the top, bottom, left and right boundaries of the foreground pixels in the entire input sequence are found, and this boundary is used as the bounding box for the entire video. Then the resize method is applied to normalize the temporal and spatial dimensions of all the motion videos to the same size. Finally, the depth value is normalized to the range of 0-1.

4.2. Experimental Results on MSR-Action3D Dataset

During the experiment, we store the features of the depth sequence from the 3D spatial-temporal convolutional neural network, the second 3D convolutional neural network and the 3D maximal pooling layer as corresponding 3D convolutional features for later SVM classification usage.

Similarly, we also use the depth motion maps corresponding to three subsets AS1, AS2 and AS3 and feed them into another 3D space-time convolution neural network for extracting the 3D convolution features.

In the fusion phase, we train the model with 3D convolutional features from the depth map sequence, depth motion map and 3D human skeleton manifold representation features corresponding to each subset respectively using linear SVM. In the test phase, three corresponding probability vectors are obtained as the SVM output. Assuming that there are \( n \) action categories, the probability vector is an \( n \)-dimensional vector and can be expressed as \( P_i = [p_{i1}, p_{i2}, \ldots, p_{in}] \), where \( p_{in} \) denotes the probability that this action is belonging to the \( n \)th action class for the \( i \)th model \( (i = 1, 2, 3) \). Given the fusion weights \( w_i \) \( (i = 1, 2, 3) \), the recognized action category index can be calculated from the following equation:

\[
    c_i = \max_j \left( \sum_{i=1}^{3} w_ip_{ij} \right) \quad (j = 1, 2, \ldots, n)
\]

In MSR-Action3D dataset, we use the same setup as the method of [43]. The results of our proposed methods and other state-of-the-art methods are shown in Table 2. As shown in this table, it can be concluded that our proposed method achieves the best results on all three subsets AS1, AS2 and AS3. This shows that our fusion method not only learns the structure and motion information through the 3D convolutional neural network, but also the details of the motion through the skeleton sequence.

|                | Method of [43] | Method of [44] | Method of [31] | Our Method |
|----------------|----------------|----------------|----------------|------------|
| AS1            | 72.9%          | 87.98%         | 74.5%          | 95.89%     |
| AS2            | 71.9%          | 85.48%         | 46.1%          | 87.68%     |
| AS3            | 79.2%          | 63.45%         | 96.4%          | 97.32%     |
| Whole dataset  | 74.7%          | 78.97%         | 82.3%          | 94.15%     |

Figures 3–5 show the confusion matrix of our proposed methods in AS1, AS2 and AS3 respectively. In Figure 3, it can be found that the recognition accuracy of the actions “hammer” and “forward punch” is relatively low. This is because these two actions are very similar to each other. Similar reasons also apply to the actions “pick and throw” and “bend”. Similar problem could also be observed in Figure 4 for actions “hand catch” and “high arm wave”. Observing the confusion matrix in Figure 5, we can find out that the recognition accuracy is better than the first two subsets since the actions in AS3 are complex and not similar, which further proves that our approach can extract the entire structure and motion information as well as the fine features of motion.
Table 3 compares the average classification accuracy of our proposed method on the MSR-Action3D dataset against other state-of-the-art methods. From Table 3, it can be concluded that the average classification accuracy of our proposed method is higher than all other competitors. This reflects the superiority of our proposed method, which can extract more space-time characteristics to enhance the discriminative power.
Table 3. Comparisons of the average classification accuracy on MSR-Action3D dataset.

| Methods                | Average Classification Accuracy |
|------------------------|--------------------------------|
| Ref. [6]               | 82.22%                         |
| Ref. [41]              | 88.73%                         |
| Ref. [45]              | 93.09%                         |
| Ref. [36]              | 91.70%                         |
| Ref. [46]              | 90.53%                         |
| Ref. [47]              | 93.63%                         |
| Our proposed method    | 94.15%                         |

4.3. Experimental Results on UTD-MHAD Dataset

UTD-MHAD dataset contains a total of 27 action categories. We classify a subset with 7 action categories. If only the 3D space-time convolutional neural network is applied to the depth map sequence, the classification accuracy is 85.7%. While using the fusion framework proposed in this paper, the classification accuracy jumps to 95.34%, which is obviously improved compared with the result without fusion. This verifies the effectiveness of our fusion framework. Our method can extract more discriminative features from both global structure/motion information and local detail features from the actions.

The confusion matrix of our proposed method is shown in Figure 6. From Figure 6 it can be concluded that our proposed method correctly identifies most of the actions. Some similar actions such as “right arm swipe to the left” and “right arm swipe to the right” can still be well separated. This again verifies the powerfulness of our proposed feature fusion scheme.

4.4. Experimental Results on UTKinect-Action3D Dataset

Firstly, we only apply 3D spatial-temporal convolutional neural networks to the depth map sequence. Unfortunately, the frame number of the videos in this dataset varies widely. For example, the seventh action “push” only contains 6 frames, while the fifth action “handling” contains about 100 frames. Therefore, after the normalization of the frame number to the same number, say 38, there will be a certain amount of information loss for the sequences with large number of frames. If only the 3D space-time convolutional neural network is applied to the depth map sequence or depth motion map, the classification accuracies are only 75.00% and 68.89% respectively. The classification accuracy of our proposed fusion based approach can achieve 97.29%. Although 3D skeleton manifold representation can get a satisfactory result, our method still makes some improvements over it, which shows the effectiveness of our fusion method.
The confusion matrix of our proposed method is shown in Figure 7. Our method can successfully identify most of the actions. The action “throw” has a relatively low recognition accuracy since the action is similar to other actions, and the motion information is not extracted completely.

![Confusion matrix](image)

**Figure 7.** Confusion matrix of our proposed method in UTKinect-Action3D dataset.

We compare the average classification accuracy of our method against other state-of-the-art methods. The comparison results are shown in Table 4. Our proposed method again achieves the highest classification accuracy.

| Methods          | Average Classification Accuracy |
|------------------|--------------------------------|
| Ref. [48]        | 96.00%                         |
| Ref. [35]        | 91.50%                         |
| Ref. [45]        | 90.91%                         |
| Ref. [19]        | 90.92%                         |
| Our proposed method | **97.29%**                      |

**Table 4.** Comparisons of the average classification accuracy on UTKinect-Action3D dataset.

5. Discussion

In this paper, we propose a fusion based action recognition framework, which mainly consists of three components: 3D space-time convolutional neural network, human skeleton manifold representation and classifier fusion. Our fusion based framework combines global structure information, motion information and details information of the action together. It automatically learns advanced features from the input data sources with the help of space-time convolutional neural networks so that it can easily adapt to different working scenarios. A simple multiply-score fusion technique is proposed in order to fully utilize the decision results from different aspects. Experimental results on three popular public benchmark datasets proved that our proposed method is the best among the current state-of-the-art methods. Although our method introduced an additional fusion step, this involves a simple SVM inference step and scores multiplications. The computational cost overhead of the additional fusion step is in the order of tens of milliseconds, which is relatively low. Furthermore, three streams in our proposed method could run in parallel so that it is still quite efficient to be executed on modern multi-core CPUs. The parallel processing time of feature learning step for three streams is about 200ms on a PC equipped with Intel 3.4GHz i7-4770 CPU.

Our method is suitable for human-computer interaction applications. For example, our proposed method could be used in controlling TV or video games in the living room. The drawbacks of our proposed method is that our method could only work in batch mode, which is to say the depth video
sequence has to be collected before sending it to the network. In the future, we plan to extend our work for online action recognition applications.

**Author Contributions:** Y.S. and Q.W. conceived and designed the experiments; C.Z., M.C. and J.Z. analyzed the data; C.Z. and Y.S. wrote the paper.

**Funding:** This research was partially funded by National Natural Science Foundation of China under contract number 61501451 and 61572020.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

**References**

1. Rahmani, H.; Mian, A.; Shah, M. Learning a deep model for human action recognition from novel viewpoints. *IEEE Trans. Pattern Anal. Mach. Intell.* 2018, 40, 667–681. [CrossRef] [PubMed]

2. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Learning clip representations for skeleton-based 3D action recognition. *IEEE Trans. Image Proc.* 2018, 27, 2842–2855. [CrossRef] [PubMed]

3. Asif, U.; Bennamoun, M.; Sohel, F. RGB-D object recognition and grasp detection using hierarchical cascaded forests. *IEEE Trans. Robot.* 2017, 33, 547–564. [CrossRef]

4. Sanchez-Riera, J.; Srinivasa, K.; Hua, K.; Cheng, W.; Hospain, M.; Alhamid, M. Robust RGB-D hand tracking using deep learning priors. *IEEE Trans. Circ. Syst. Video Technol.* 2018, 28, 2289–2301. [CrossRef]

5. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011.

6. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.

7. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 2014, 36, 914–927. [CrossRef] [PubMed]

8. Althloothi, S.; Mahoor, M.; Zhang, X.; Velyes, R. Human activity recognition using multi-features and multiple kernel learning. *Pattern Recog.* 2014, 47, 1800–1812. [CrossRef]

9. Chaaraoui, A.; Padiallopez, J.; Florezrevuelta, F. Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In Proceedings of the International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013.

10. Lv, F.; Nevatia, R. Recognition and segmentation of 3D human action using HMM and multi-class Adaboost. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006.

11. Azary, S.; Savakis, A. Grassmannian sparse representations and motion depth surfaces for 3D action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 25–27 June 2013.

12. Anirudh, R.; Turaga, P.; Su, J.; Srivastava, A. Elastic functional coding of human actions: From vector-fields to latent variables. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.

13. Luo, J.; Wang, W.; Qi, H. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In Proceedings of the International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013.

14. Gong, D.; Medioni, G. Dynamic manifold warping for view invariant action recognition. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.

15. Slama, R.; Wannous, H.; Daoudi, M.; Srivastava, A. Accurate 3D action recognition using learning on the Grassmann manifold. *Pattern Recognit.* 2015, 48, 556–567. [CrossRef]

16. Zhu, Y.; Chen, W.; Guo, G. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image Vis. Comput.* 2014, 32, 453–464. [CrossRef]

17. Chen, C.; Liu, M.; Liu, H.; Zhang, B.; Han, J.; Kehtarnavaz, N. Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition. *IEEE Access* 2017, 5, 22590–22604. [CrossRef]
18. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, CA, USA, 3–6 December 2012.

19. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P. Deep convolutional neural networks for action recognition using depth map sequences. arXiv, 2015; arXiv:1501.04686.

20. Gong, D.; Medioni, G.; Zhao, X. Structured time series analysis for human action segmentation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2014, 36, 1414–1427. [CrossRef]

21. Lefebvre, G.; Berlemont, S.; Mamalet, F.; Garcia, C. BLSTM-RNN based 3D gesture classification. In Proceedings of the International Conference on Artificial Neural Networks, Sofia, Bulgaria, 10–13 September 2013.

22. Serre, T.; Wolf, L.; Bileschi, S.; Riesenhuber, M.; Poggio, T. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 411–426. [CrossRef] [PubMed]

23. Lefebvre, G.; Berlemont, S.; Mamalet, F.; Garcia, C. BLSTM-RNN based 3D gesture classification. In Proceedings of the International Conference on Artificial Neural Networks, Sofia, Bulgaria, 10–13 September 2013.

24. Serre, T.; Wolf, L.; Bileschi, S.; Riesenhuber, M.; Poggio, T. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 411–426. [CrossRef] [PubMed]

25. Jhuang, H.; Serre, T.; Wolf, L.; Poggio, T. A biologically inspired system for action recognition. In Proceedings of the International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.

26. Jhuang, H.; Serre, T.; Wolf, L.; Poggio, T. A biologically inspired system for action recognition. In Proceedings of the International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.

27. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. C3D: Generic features for video analysis. arXiv 2014, arXiv:1412.0767v1.

28. Wang, P.; Cao, Y.; Shen, C.; Liu, L.; Shen, H. Temporal pyramid pooling-based convolutional neural network for action recognition. IEEE Trans. Circ. Syst. Video Technol. 2017, 27, 2613–2622. [CrossRef]

29. Yang, X.; Tian, Y. EigenJoints-based action recognition using naïve-bayes-nearest-neighbor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.

30. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. C3D: Generic features for video analysis. arXiv 2014, arXiv:1412.0767v1.

31. Wang, P.; Cao, Y.; Shen, C.; Liu, L.; Shen, H. Temporal pyramid pooling-based convolutional neural network for action recognition. IEEE Trans. Circ. Syst. Video Technol. 2017, 27, 2613–2622. [CrossRef]

32. Wang, C.; Wang, Y.; Yuille, A. An approach to pose-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.

33. Sheikh, Y.; Sheikh, M.; Shah, M. Exploring the space of a human action. In Proceedings of the International Conference on Computer Vision, Beijing, China, 17–20 October 2005.

34. Chaudhry, R.; Ofli, F.; Kurillo, G.; Bajcsy, R.; Vidal, R. Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–24 June 2013.

35. Hussein, M.; Torki, M.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 5–9 August 2013.

36. Zanfir, M.; Leordeanu, M.; Sminchisescu, C. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In Proceedings of the International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013.

37. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3D skeletons as points in a Lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.

38. Sharaf, A.; Torki, M.; Hussein, M.; El-Saban, M. Real-time multiscale action detection from 3D skeleton data. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015.
39. Devanne, M.; Wannous, H.; Berretti, P.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3D human action recognition by shape analysis of motion trajectories on Riemannian manifold. *IEEE Trans. Cybern.* 2015, 45, 1340–1352. [CrossRef] [PubMed]

40. Hunter, D.; Yu, H.; Pukish, M.; Kolbusz, J.; Wilamowski, B. Selection of proper neural network sizes and architectures: A comparative study. *IEEE Trans. Ind. Inform.* 2012, 8, 228–240. [CrossRef]

41. Yang, X.; Zhang, C.; Tian, Y. Recognizing actions using depth motion map-based histograms of oriented gradients. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012.

42. UTD-MHAD. Available online: http://www.utdallas.edu/~kehtar/UTD-MHAD.html (accessed on 12 December 2018).

43. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3D points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010.

44. Xia, L.; Chen, C.; Aggarwal, J. View invariant human action recognition using histograms of 3D joints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.

45. Yang, X.; Tian, Y. Super normal vector for activity recognition using depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.

46. Yang, Y.; Deng, C.; Tao, D.; Zhang, S.; Liu, W.; Gao, X. Latent max-margin multitask learning with skelets for 3-D action recognition. *IEEE Trans. Cybern.* 2017, 47, 439–448. [CrossRef] [PubMed]

47. Yang, Y.; Deng, C.; Gao, S.; Liu, W.; Tao, D.; Gao, X. Discriminative multi-instance multitask learning for 3D action recognition. *IEEE Trans. Multimedia* 2017, 19, 519–529. [CrossRef]

48. Liu, Z.; Zhang, C.; Tian, Y. 3D-based deep convolutional neural network for action recognition with depth sequences. *Image Vis. Comput.* 2016, 55, 93–100. [CrossRef]