sensors

MDPI

Article

Integrally Cooperative Spatio-Temporal Feature Representation of Motion Joints for Action Recognition

Xin Chao, Zhenjie Hou *, Jiuzhen Liang and Tianjin Yang

School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, China; chaoxin941203@163.com (X.C.); jzliang@cczu.edu.cn (J.L.); yangtianjin128@163.com (T.Y.)

* Correspondence: houzj@cczu.edu.cn

Received: 12 August 2020; Accepted: 9 September 2020; Published: 11 September 2020

Abstract: In contemporary research on human action recognition, most methods separately consider the movement features of each joint. However, they ignore that human action is a result of integrally cooperative movement of each joint. Regarding the problem, this paper proposes an action feature representation, called Motion Collaborative Spatio-Temporal Vector (MCSTV) and Motion Spatio-Temporal Map (MSTM). MCSTV comprehensively considers the integral and cooperative between the motion joints. MCSTV weighted accumulates limbs' motion vector to form a new vector to account for the movement features of human action. To describe the action more comprehensively and accurately, we extract key motion energy by key information extraction based on inter-frame energy fluctuation, project the energy to three orthogonal axes and stitch them in temporal series to construct the MSTM. To combine the advantages of MSTM and MCSTV, we propose Multi-Target Subspace Learning (MTSL). MTSL projects MSTM and MCSTV into a common subspace and makes them complement each other. The results on MSR-Action3D and UTD-MHAD show that our method has higher recognition accuracy than most existing human action recognition algorithms.

Keywords: human action recognition; Motion Collaborative Spatio-Temporal Vector; Motion Spatio-Temporal Map; Multi-Target Subspace Learning; key information extraction based on inter-frame energy fluctuation

1. Introduction

Human action recognition [1] is a research hotspot in the field of artificial intelligence and pattern recognition. The research achievements have been used in many aspects of life [2], such as human-computer interaction, biometrics, health monitoring, video surveillance systems, somatosensory game, robotics, etc. [3].

Due to the development of lower-cost depth sensors, deep cameras have been widely used in action recognition. Compared with traditional red, green and blue (RGB) cameras, the depth camera is not sensitive to lighting conditions [4]. It is easy to distinguish the background and foreground, and provides human depth data. In addition, human skeletal information can also be obtained from the depth map.

So far, many scholars have used depth sequences for human action recognition research. Li et al. [5] selected representative 3D points to depict human action. Oreifej et al. [6] proposed histogram of oriented 4d normals (HON4D) to capture the structural information of action. Yang et al. [7] proposed depth motion map (DMM) to accumulate the motion energy of neighboring moments.

Since the skeletal information can directly describe the movement of human joints, many scholars have begun to use skeleton sequences to recognize human action. Xia et al. [8] proposed histograms of 3D joints (HOJ3D) as a compact representation of human action. Vemulapalli et al. [9] represented

human actions as curves that contain skeletal action sequences. They modeled the 3D geometric correlations among different body parts by using rotations and translations. Luvizon et al. [10] proposed a robust method based on vector of locally aggregated descriptors (VLAD) algorithm and clustering library, which extracts spatiotemporal local feature sets from joint subgroups to represent human action.

In the field of action recognition, the use of single modal data is one-sided, and it is necessary to integrate different modals data for comprehensive decision-making. Many fusion methods have been successfully applied to action recognition, among which feature fusion is most extensive. Canonical Correlation Analysis (CCA) [11] and Discriminant Correlation Analysis (DCA) [12] are commonly used feature fusion methods.

Although significant progress has been made in human action recognition, there are still shortcomings. Currently, many studies only independently perform feature extraction [13] and recognition of some motive joints. However, they did not consider the comprehensiveness, integration and collaboration between the joints. For example, in throwing, the main movement part of the body is right hand. If we only consider the movement features of right hand, this action is likely to be recognized as a waving. If we combine the movements of the left hand, left and right legs of the body with those of the right hand, then the possibility of this action being recognized as a throwing is greatly enhanced. Regarding the problem, this paper proposes an action feature representation algorithm that considers the integrally cooperative movement features of human action, called Motion Collaborative Spatio-Temporal Vector (MCSTV) and Motion Spatio-Temporal Map (MSTM). MCSTV reflects the integrally cooperative of human action through weighted accumulating limbs' motion vector. MSTM can accurately describe the spatial structure [14] and temporal information [15] of actions, we extract key motion energy by key information extraction based on inter-frame energy fluctuation, and project the key energy of body on three orthogonal axes and stitched according to temporal series to form three-axis MSTMs. To give full play to the advantages of MSTM and MCSTV, we propose Multi-Target Subspace Learning (MTSL). MTSL projects MSTM and MCSTV into a common subspace, alienates the inter-class distance of samples of different categories and reduces the dimension of projection target area by constructing multiple projection target centers of each sample of the same category. The workflow illustration of our method is shown in Figure 1.

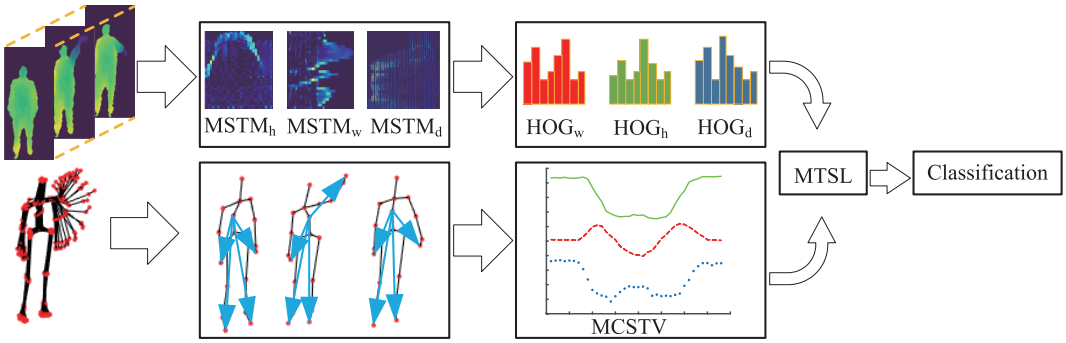

Figure 1. Workflow illustration of our method.

In this paper, we focus on the challenging problem of action recognition. Our contributions are as follows:

- Motion Collaborative Spatio-Temporal Vector is proposed, which is a feature representation method that comprehensively considers the integral and cooperative of human action.
- Motion Spatio-Temporal Map is proposed, which is a feature representation method that fully preserves the spatial structure and temporal information of human action.
- Key information extraction based on inter-frame energy fluctuation is proposed, which is a method that extracts key information in motion energy.
- Multi-Target Subspace Learning is proposed, which is used to fuse MCSTV and MSTM.

This paper is organized as follows. In Section 2, the related work is briefly reviewed. In Section 3, method is detailly described. The results of experimental and discussions are presented in Section 4. Finally, Section 5 summarizes the paper.

2. Related Work

2.1. Action Recognition Based on Skeleton Approach

Human action recognition based on skeleton approach is a research hotspot. Lv et al. [16] decomposed the high dimensional 3D joint space into a set of feature space. Then, the hidden Markov model (HMM) combined with multi-class AdaBoost (Adabbs.m2) algorithm and dynamic programming algorithm are used to segment and recognize the actions. Hussein et al. [17] presented an action recognition algorithm from 3D skeleton sequences extracted from depth data. This method uses the covariance matrix for skeleton joint locations over time as a discriminative descriptor for a sequence. Zanfir et al. [18] proposed a non-parametric Moving Pose (MP) framework for human action recognition. They captured both pose information and the speed and acceleration of human joints and finally used a modified K-Nearest Neighbor (KNN) classifier in conjunction with the descriptors to classify human actions. However, these methods did not consider the comprehensiveness, integration and collaboration between the joints.

2.2. Motion Energy Image and Motion History Image

In the early stage of human action recognition, various methods were used to extract features from color video collected by RGB cameras to complete recognition. Bobick et al. [19] proposed Motion Energy Image (MEI) and Motion History Image (MHI). The MEI algorithm needs to extract the foreground area of movement. Then the foreground region is binarized to obtain the binary image sequence of the action. Finally, the union of binary image sequences is taken to obtain MEI of the action. The calculation of MEI is expressed as:

$$\mathrm{MEI}_{\delta}(x,y,t) = \bigcup_{i=0}^{\delta-1} B(x,y,t-i) \tag{1}$$

where $\mathrm{MEI}_{\delta}(x,y,t)$ represents that MEI is generated by δ images at the t-th frame in the video sequence. $B(x,y,t)$ represents the t-th frame of the binary image sequence. x and y respectively represent the height and width of a pixel point on the image. t represents the sequence number of a frame in the video sequence.

The MEI describes the largest contour of actions. It loses some motion information inside the contour and cannot express the temporal information of actions.

Unlike MEI, MHI is a grayscale image. The pixel intensity is a function of the temporal history of motion at that point. A simple replacement and decay operator can be used:

$$\mathrm{MHI}_{\tau}(x,y,t) = \begin{cases} \tau & \text{if } B(x,y,t)=1 \\ \max\left(0, \mathrm{MHI}_{\tau}(x,y,t-1)-1\right) & \text{otherwise} \end{cases} \tag{2}$$

where τ is the initial brightness; $\mathrm{MHI}_{\tau}(x,y,t)$ is the MHI generated by the t frames of the video sequence.

Compared with MEI, MHI retains part temporal information of actions through brightness attenuation. However, MHI also cannot fully express the spatial structure.

2.3. Action Recognition Based on Depth Approach

With the development of depth sensors, scholars began to use depth sequence to study human action recognition. Numerous scholars [20] use depth motion map (DMM) for action recognition

research. Each frame in the depth sequence is projected onto three orthogonal cartesian planes to form three 2D projection maps according to front, side, and top views, denoted map_f, map_s, map_t. The motion energy of two consecutive projection images is respectively accumulated to form the DMM of three views. The calculation of DMM is expressed as:

$$\text{DMM}_v = \sum_{i=1}^{N} \left(\left| \text{map}_v^{i+1} - \text{map}_v^{i} \right| > \varepsilon \right) \tag{3}$$

where $v \in \{\text{f}, \text{s}, \text{t}\}$ represents the projection view, f is front view, s is side view, t is top view. DMM_v is the DMM of view v. $\text{map}_v^{i+1} - \text{map}_v^{i}$ denotes the image difference between the i-th frame and the $i+1$-th frame, namely the motion energy between the i-th frame and the $i+1$-th frame. N represents the number of frames of the depth sequence. ε indicates a difference threshold.

Compared with MEI, DMM fully reflects the depth information of actions. However, DMM also cannot express the temporal information.

Yang et al. [7] computed Histograms of Oriented Gradients (HOG) from DMM as the representation of an action sequence. Chen et al. [20] extracted Local Binary Pattern (LBP) from DMM. Chen et al. [21] extracted Gradient Local Auto-Correlations (GLAC) from DMM. Zhang et al. [22] presented 3D histograms of texture (3DHoTs) to extract discriminant features from depth video sequences. Wang et al. [23] extracted random occupancy pattern (ROP) features from depth video sequences and use a sparse coding approach to encode these features. Xia et al. [24] built depth cuboid similarity features (DCSF) around the local spatio-temporal interest points (STIPs) extracted from depth video sequences to describe local 3D depth cuboids. Vieira et al. [25] presented Space-Time Occupancy Patterns (STOP) to preserve spatial and temporal information between space-time cells.

2.4. Feature Fusion

Feature fusion can make single features complement each other to improve recognition accuracy. Many achievements have been made in this field. For example, Hardoon et al. [11] proposed Canonical Correlation Analysis (CCA) which maximizes the correlation between two different features. Zhang et al. [26] proposed group sparse canonical correlation analysis to preserve the group sparse characteristics of data within each set in addition to maximize the global covariance. Haghighat et al. [12] proposed DCA that performs feature fusion by maximizing the pairwise correlations across the two feature sets and eliminating the inter-class correlations and restricting the correlations to be in the classes. Wang et al. [27] proposed Joint Feature Selection and Subspace Learning (JFSSL) for for cross-modal retrieval, in which multimodal features are projected into a common subspace by learning projection matrices. However, these methods generally describe the correlation between different features and do not consider the correlation between the samples of different categories.

3. Method

As shown in Figure 1, we extract MSCTV from skeleton sequences, and extract MSTM from depth video sequences. Next, we further extract the Histogram of Oriented Gridient (HOG) feature of MSTM. Then, we fuse MCSTV and MSTM through MTSL, and use the fusion features to complete human action recognition research.

3.1. Motion Collaborative Spatio-Temporal Vector

Human action is the fact of integrally cooperative movement of all joints, not the fact of individual movement of some joints. Therefore, this paper proposes Motion Collaborative Spatio-Temporal Vector that considers the integrally cooperative movement of limbs' joints.

Most actions are the result of multiple joints moving integrally. In this paper, the scattered and separate motion joints' information is spliced by vector superposition to form a comprehensive vector

that can comprehensively describe action and highlight the integral and cooperative of movement. The comprehensive vector is called MCSTV, the basic principle of MCSTV is shown in Figure 2. Where $\overrightarrow{SLH}$ represents the motion vector of left upper limb, $\overrightarrow{SRH}$ represents the motion vector of right upper limb, $\overrightarrow{SLF}$ represents the motion vector of left lower limb, $\overrightarrow{SRF}$ represents the motion vector of right lower limb, and $\overrightarrow{MCST}_o = \overrightarrow{SLH} + \overrightarrow{SRH} + \overrightarrow{SLF} + \overrightarrow{SRF}$, $\overrightarrow{MCST}_o$ represents the original MCSTV. Due to MCSTV is stitched by the limbs' motion vector, MCSTV can reflect the fact of the comprehensive effect of multiple motion vectors to a certain degree. The human skeleton is shown in Figure 3a. The skeletal frame of the high wave is shown in Figure 3b.

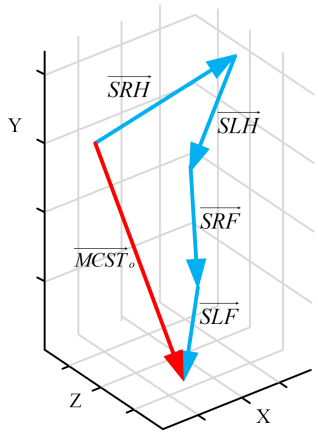

Figure 2. The basic idea of MCSTV (Contribution degree is not considered).

As shown in Figure 3b, we select the motion vector from the spine to the left hand to represent the motion of left upper limb, that from the spine to the right hand to represent the motion of right upper limb, that from the spine to the left foot to represent the motion of left lower limb, and that from the spine to the right foot to represent the motion of right lower limb, respectively, denoted $\overrightarrow{SLH}$, $\overrightarrow{SRH}$, $\overrightarrow{SLF}$, $\overrightarrow{SRF}$.

For different actions, the contribution degree of each joint is different. As showed in Figure 2, the motion vector of the limbs are directly accumulated. Due to the degree of limb is different, if limbs' motion vector are directly accumulated, the action cannot be accurately described. We must obtain the contribution degree of each limb. However, these motion vectors are all three-dimensional vectors, and the change of the vector in space is the result of the change of the vector in multi-view. Therefore, these motion vectors need to be projected onto three orthogonal cartesian planes xoy, yoz and xoz. The offset of $\overrightarrow{SLH}$ on each plane is expressed as:

$$\begin{aligned}
\theta_{SLH_{xy}}(i) &= \arccos\left(\frac{\overrightarrow{SLH}_{xy}(i+1)\times\overrightarrow{SLH}_{xy}(i)}{\left|\overrightarrow{SLH}_{xy}(i+1)\right|\left|\overrightarrow{SLH}_{xy}(i)\right|}\right)\\
\theta_{SLH_{yz}}(i) &= \arccos\left(\frac{\overrightarrow{SLH}_{yz}(i+1)\times\overrightarrow{SLH}_{yz}(i)}{\left|\overrightarrow{SLH}_{yz}(i+1)\right|\left|\overrightarrow{SLH}_{yz}(i)\right|}\right)\\
\theta_{SLH_{xz}}(i) &= \arccos\left(\frac{\overrightarrow{SLH}_{xz}(i+1)\times\overrightarrow{SLH}_{xz}(i)}{\left|\overrightarrow{SLH}_{xz}(i+1)\right|\left|\overrightarrow{SLH}_{xz}(i)\right|}\right)
\end{aligned} \tag{4}$$

where $\overrightarrow{SLH}_{xy}(i)$, $\overrightarrow{SLH}_{yz}(i)$ and $\overrightarrow{SLH}_{xz}(i)$ respectively represent the projection of the i-th frame of $\overrightarrow{SLH}$ on xoy, yoz and xoz. $\theta_{SLH_{xy}}(i)$, $\theta_{SLH_{yz}}(i)$ and $\theta_{SLH_{xz}}(i)$ respectively represent the offsets of the i-th frame of $\overrightarrow{SLH}$ on xoy, yoz and xoz.

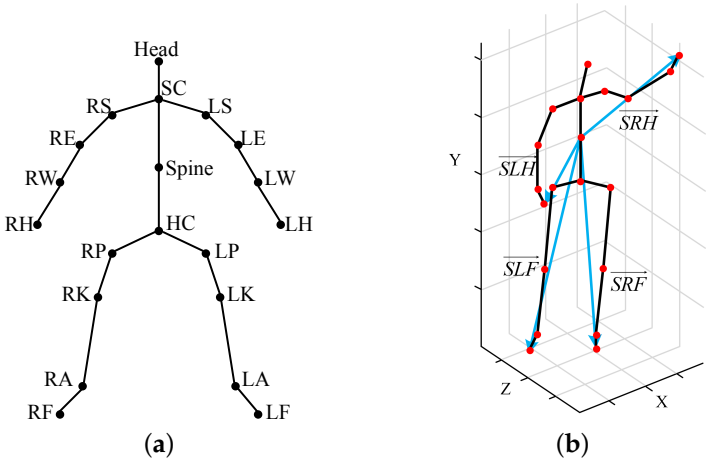

Figure 3. Human skeleton. (**a**) Skeleton; (**b**) The skeletal frame of the high wave.

The offset of $\overrightarrow{SLH}$ of each frame is the sum of offset on three orthogonal planes. The offset of each frame of $\overrightarrow{SLH}$ is expressed as:

$$\theta_{SLH}(i) = \theta_{SLH_{xy}}(i) + \theta_{SLH_{yz}}(i) + \theta_{SLH_{xz}}(i) \tag{5}$$

where $\theta_{SLH}(i)$ represents the offsets of the i-th frame of $\overrightarrow{SLH}$.

Each action consists of N frames, so the total offset of $\overrightarrow{SLH}$ is the sum of the offset of each frame. The offset of $\overrightarrow{SLH}$ is expressed as:

$$sum_{SLH} = \sum_{i=1}^{N} \theta_{SLH}(i) \tag{6}$$

where sum_{SLH} represents the total offset of $\overrightarrow{SLH}$.

Similarly, according to Equations (4)–(6), the total offset of $\overrightarrow{SRH}$ is obtained as sum_{SRH}, the total offset of $\overrightarrow{SLF}$ as sum_{SLF} and the total offset of $\overrightarrow{SRH}$ as sum_{SRF}.

The contribution degree of each limb is the ratio of the offset of the limb and the total offset of all limbs. The contribution degree of each limb is expressed as:

$$\begin{aligned} W_{SLH} &= \frac{sum_{SLH}}{sum_{SLH}+sum_{SRH}+sum_{SLF}+sum_{SRF}} \\ W_{SRH} &= \frac{sum_{SRH}}{sum_{SLH}+sum_{SRH}+sum_{SLF}+sum_{SRF}} \\ W_{SLF} &= \frac{sum_{SLF}}{sum_{SLH}+sum_{SRH}+sum_{SLF}+sum_{SRF}} \\ W_{SRF} &= \frac{sum_{SRF}}{sum_{SLH}+sum_{SRH}+sum_{SLF}+sum_{SRF}} \end{aligned} \tag{7}$$

where W_{SLH}, W_{SRH}, W_{SLF} and W_{SRF} respectively represent the contribution degree of $\overrightarrow{SLH}$, $\overrightarrow{SRH}$, $\overrightarrow{SLF}$ and $\overrightarrow{SRF}$.

Finally, the contribution degree of each limb was used to constrain motion vector of each frame, and motion vector of each limb were weighted accumulated to form MCSTV. The calculation of MCSTV is expressed as:

$$\overrightarrow{MCST}_w(i) = \overrightarrow{SLH}(i) \times W_{SLH} + \overrightarrow{SRH}(i) \times W_{SRH} + \overrightarrow{SLF}(i) \times W_{SLF} + \overrightarrow{SRH}(i) \times W_{SRH} \tag{8}$$

where $\overrightarrow{MCST}_w(i)$ represents the i-th frame of $\overrightarrow{MCST}_w$, $\overrightarrow{MCST}_w$ represents the MCSTV that motion vector of each limb were weighted accumulated.

MCSTV is obtained by weighted accumulation of motion vector of each limb is shown in Figure 4. As can be seen from the figure, after weighted accumulation, the MCSTV of this action is dominated by $\overrightarrow{SRH}$. Compared with the direct accumulation of motion vector in Figure 2, this method more directly reflects the major motion joints of the action.

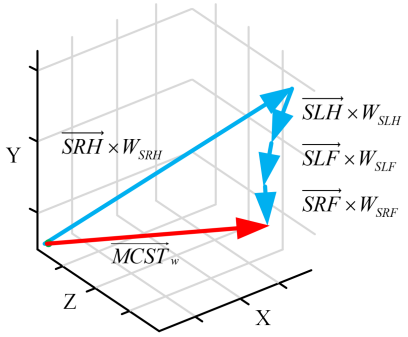

Figure 4. The final MCSTV.

3.2. Motion Spatio-Temporal Map

To describe the action information more comprehensively and accurately, this paper proposes a feature representation algorithm, called Motion Spatio-Temporal Map (MSTM). MSTM can completely express the spatial structure and temporal information. This algorithm calculates the difference between adjacent frames of depth sequence to obtain the motion energy. Next, the key information is extracted from the motion energy by key information extraction based on inter-frame energy fluctuation. Then, the key energy is projected to three orthogonal axes to obtain the motion energy list of the three orthogonal axes. Finally, the motion energy list is spliced in temporal series to form MSTM. The flow illustration of MSTM is shown in Figure 5.

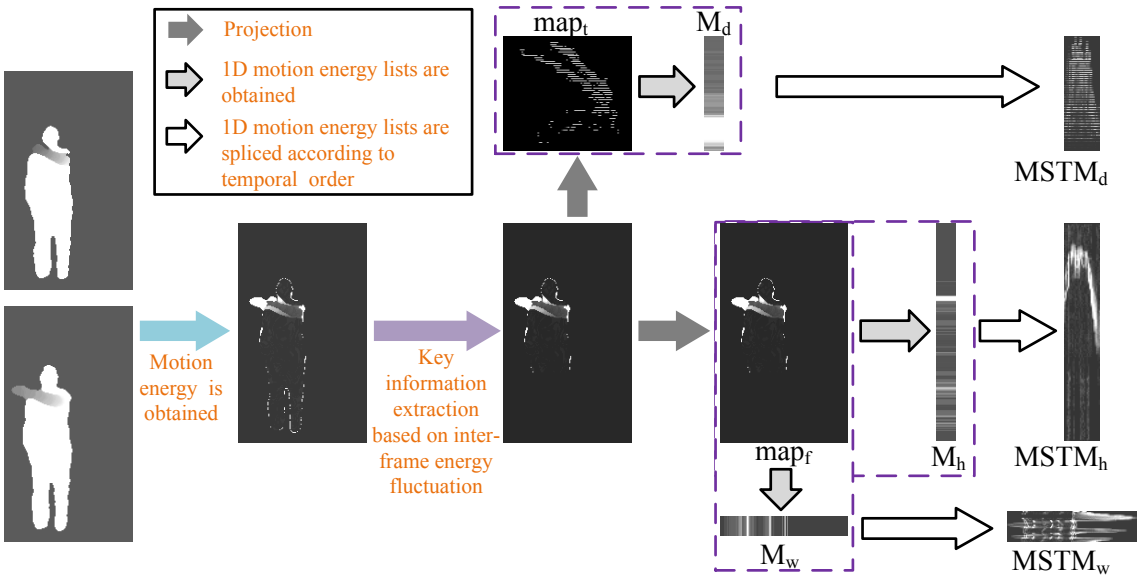

Figure 5. The flow illustration of MSTM.

As shown in Figure 5, the motion energy of the action is obtained through the difference operation between two adjacent frames of the depth sequence. The motion energy is expressed as:

$$E_k(x,y,z) = |I_{k+1}(x,y,z) - I_k(x,y,z)| \tag{9}$$

where $I_k(x, y, z)$ and $I_{k+1}(x, y, z)$ respectively represent the body energy of the k-th and k+1-th moment, i.e., the k-th and k+1-th frame of depth video sequence. $E_k(x, y, z)$ represents the motion energy of the k-th moment.

Due to the habitual sloshing of some joints, there is a lot of redundancy in the motion energy obtained by Equation (9). Regarding the problem, we propose an energy selection algorithm, i.e., key information extraction based on inter-frame energy fluctuation. We use this algorithm to remove the redundant of the motion energy at each moment. The main idea of this algorithm is to divide the human body into four areas according to the height range and width range at the initial moment of actions. Then, we calculate the proportion of the motion energy of each region in the whole body, and select a certain amount of motion energy as the main energy. The remaining energy is taken as redundancy and removed. The detailed steps of this algorithm are as follows:

Firstly, we calculate the height range $[h_{\min}, h_{\max}]$ and width range $[w_{\min}, w_{\max}]$ of human activity area at the initial moment of actions. The body is divided into upper body and lower body according to the hip center of the body. The body is divided into left body and right body according to the symmetry. Finally, the body is divided into four regions: left upper body (LU), right upper body (RU), left lower body (LL), and right lower body (RL). The motion energy of the body is the sum of the motion energy of four regions. The division of human activity areas is shown in Figure 6.

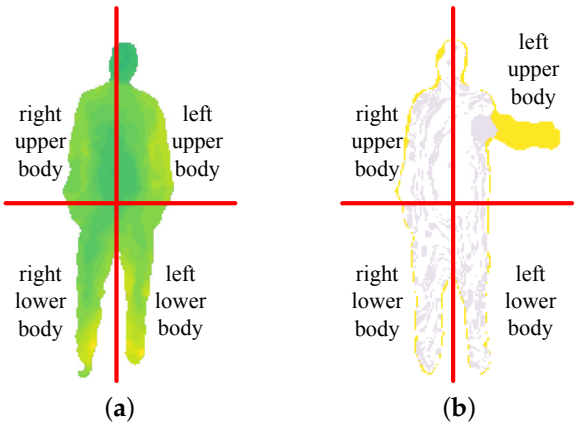

Figure 6. The division of human activity areas. (**a**) The division of human activity areas at the beginning of the action; (**b**) The division of human activity areas when the wave.

Next, we calculate the motion energy proportion of each region in the whole body, the energy proportion can be expressed as:

$$\eta_\psi = \frac{\sum_{x=1}^{H_\psi} \sum_{y=1}^{W_\psi} E_k(x, y, z)}{\sum_{x=1}^{H} \sum_{y=1}^{W} E_k(x, y, z)} \tag{10}$$

where $\psi \in \{\mathrm{LU}, \mathrm{RU}, \mathrm{LL}, \mathrm{RL}\}$, η_ψ represent the motion energy proportion of each region. H, W respectively represent the height and width of the whole body. H_ψ, W_ψ respectively represent the height and width of each region.

Then, we rank the motion energy proportions of each region from the largest to the smallest. The maximum value is denoted as η_1, the minimum value is denoted as η_4, and $\eta_1 > \eta_2 > \eta_3 > \eta_4$. In this paper, we select ξ of the whole body energy as the key energy. The remaining energy is considered to be redundant and removed from the original motion energy. The value of ξ is determined by the experimental results and recognition accuracies in Section 4.2.1. The selection of key energy follows the following criteria.

If $\eta_1 > \xi$, then the motion energy of the corresponding region of η_1 is retained as the key energy, and the motion energy of the other three regions is considered to be redundant. If $\eta_1 < \xi$ and $\eta_1 + \eta_2 > \xi$, then the motion energy of the corresponding region of η_1 and η_2 is retained as the key

energy, and the motion energy of the other two regions is considered to be redundant. If $\eta_1 + \eta_2 < \xi$ and $\eta_1 + \eta_2 + \eta_3 > \xi$, then the motion energy of the corresponding region of η_1, η_2 and η_3 is retained as the key energy, and the motion energy of η_4 is considered to be redundant. If none of the above conditions are met, the whole body motion energy is considered to be the key energy and retained.

The key energy is projected onto three orthogonal cartesian planes to form three 2D projection maps according to front view, side view, and top view, denoted $\mathrm{map_f}$, $\mathrm{map_s}$, $\mathrm{map_t}$. The 2D projection maps are expressed as:

$$\begin{aligned} \mathrm{map_f}(x,y) &= E(x,y) \\ \mathrm{map_s}(x,z) &= E(x,z) \\ \mathrm{map_t}(y,z) &= E(y,z) \end{aligned} \tag{11}$$

where $\mathrm{map_f}(x,y)$, $\mathrm{map_s}(x,z)$ and $\mathrm{map_t}(y,z)$ respectively represent the coordinate of a pixel point on the $\mathrm{map_f}$, $\mathrm{map_s}$ and $\mathrm{map_t}$. $E(x,y)$, $E(x,z)$, and $E(y,z)$ respectively represent the value of a pixel point on the $\mathrm{map_f}$, $\mathrm{map_s}$ and $\mathrm{map_t}$.

To obtain the energy distribution of the width axis, height axis and depth axis, we select $\mathrm{map_f}$ and $\mathrm{map_t}$ to continue to project to the corresponding orthogonal axis to obtain the row sum or column sum of the 2D energy projection maps. According to width axis, height axis and depth axis, three 1D motion energy lists are generated, which are expressed as $\mathrm{M_w}$, $\mathrm{M_h}$ and $\mathrm{M_d}$ respectively. The 1D motion energy list is expressed as:

$$\begin{aligned} \mathrm{M_w}(j) &= \textstyle\sum_{x=1}^{H_m} \mathrm{map_f}(x,j) \\ \mathrm{M_h}(j) &= \textstyle\sum_{y=1}^{W_m} \mathrm{map_f}(j,y) \\ \mathrm{M_d}(j) &= \textstyle\sum_{y=1}^{W_m} \mathrm{map_t}(j,y) \end{aligned} \tag{12}$$

Where $\mathrm{M_w}(j)$, $\mathrm{M_h}(j)$, and $\mathrm{M_d}(j)$ respectively represent the j-th element of the energy list on the width axis, height axis, and depth axis. W_m and H_m respectively represent the width and height of the 2D energy projection map.

According to the temporal order, M_u are spliced to form MSTM of three axes, which are respectively represented as MSTM_w, MSTM_h and MSTM_d. For the depth sequence of N frames, the calculation of MSTM is expressed as:

$$\mathrm{MSTM}_u(k) = \mathrm{M}_u^k \tag{13}$$

where $u \in \{\mathrm{w}, \mathrm{h}, \mathrm{d}\}$, w is width axis, h is height axis and d is depth axis, M_u^k represents the 1D motion energy list of the k-th frame of the action sequence on the u axis. MSTM_u represents the MSTM on the u axis. $\mathrm{MSTM}_u(k)$ represents the k-th row of MSTM_u.

With the maximum width $W_{\max}$, minimum width $W_{\min}$, maximum height $H_{\max}$ and minimum height $H_{\max}$ of the body's activity area as the bounds, MSTM is processed with the region of interest (ROI) [28], i.e., the image is cropped and normalized.

The actions in the original depth sequence are defined as positive order actions. The positive order high throw is shown in Figure 7a. The actions that the order is contrary to the original depth sequence are defined as reverse order actions. The reverse order high throw is shown in Figure 7b. The various feature maps of the positive and reverse order high throw are shown in Figure 8.

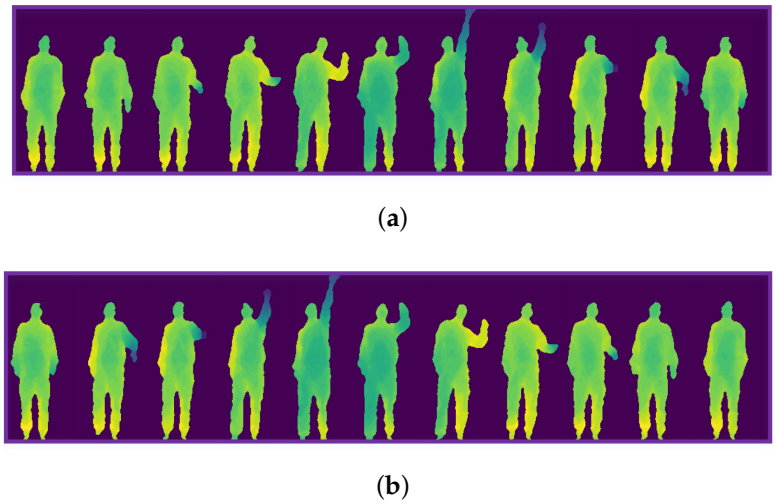

Figure 7. High throw. (**a**) The positive order of high throw; (**b**) The reverse order of high throw.

Figure 8a,b are the MSTM of positive order and reverse order high throw respectively. From left to right are the MSTM of height axis, width axis and depth axis respectively. Owing to MSTM reflects the change of the energy information on the three orthogonal axes, it preserves the spatial and temporal information completely. Positive order and reverse order actions have the same motion trajectories and opposite temporal orders, the final MSTM is symmetric along the time axis and easy to be distinguish. In contrast, Figure 8e,f respectively represent the MHI of positive order and reverse order high throw. MHI retains part of temporal information of the actions and has the ability to distinguish between positive order and reverse order actions. However, due to the coverage of the trajectory and the absence of the depth information, MHI cannot fully express the spatial information. Figure 8c,d are the MEI of positive order and reverse order high throw, respectively. Figure 8g,h are the DMM of positive order and reverse order high throw, respectively. MEI and DMM do not involve temporal information, so they cannot be distinguish. MEI does not involve the depth information, which means the spatial information is incomplete. DMM contains depth information and expresses spatial information fully.

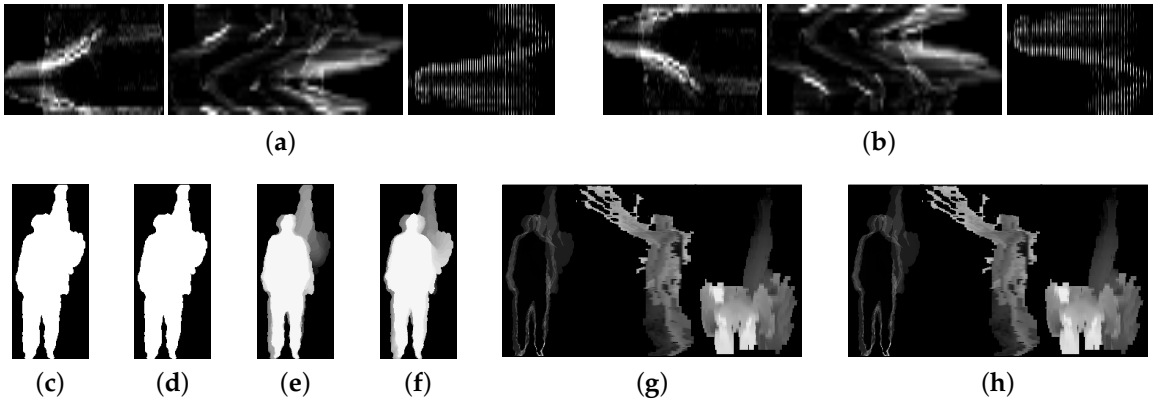

Figure 8. The various feature maps of the positive and reverse order high throw. (**a**) MSTM of positive order high throw; (**b**) MSTM of reverse order high throw; (**c**) MEI of positive order high throw; (**d**) MEI of reverse order high throw; (**e**) MHI of positive order high throw; (**f**) MHI of reverse order high throw; (**g**) DMM of positive order high throw; (**h**) DMM of reverse order high throw.

3.3. Feature Fusion

To ensure that the description of action information more accurate, fuse features are usually used in human action recognition research. Therefore, this paper fuses skeleton feature MCSTV and image

feature MSTM. It cannot only reflect the integrity and cooperativity of the action, but also express the spatial structure and temporal information more completely.

Let $\Gamma = \{\mathrm{x}_i^1, \mathrm{x}_i^2, \cdots, \mathrm{x}_i^M\}_{i=1}^N$ denote N action samples, and the i-th sample $\Gamma_i = \{\mathrm{x}_i^1, \mathrm{x}_i^2, \cdots, \mathrm{x}_i^M\}$ contains features from M different modalities, but they correspond to the same representation in a common space, denoted by y_i. Where x_i is the sample of the i-th category, y_i is the target projection center of the i-th category, modality M represents M types data.

In this paper, we propose Multi-Target Subspace Learning (MTSL) to study the common subspace of different features. The minimization problem is followed as:

$$\min_{\mathbf{U}_1,\cdots,\mathbf{U}_M} \sum_{p=1}^{M} \left\| \mathbf{X}_p^T \mathbf{U}_p - \mathbf{Y} \right\|_F^2 + \lambda_1 \sum_{p=1}^{M} \left\| \mathbf{U}_p \right\|_{21} + \lambda_2 \sum_{p=1}^{M} \sum_{c=1}^{\mathrm{L}-1} \left\| \mathbf{X}_p^T \mathbf{U}_p - G_c \right\|_F^2 \tag{14}$$

where $\mathbf{U}_p, p = 1, 2, \ldots, M$ is the projection matrix of the p-th modality. $\mathbf{X}_p$ is the sample features of the p-th modality before the projection. $\mathbf{X}_p^T \mathbf{U}_p$ is the sample features of the p-th modality after the projection. $\mathbf{Y}$ is the primary target projection matrix in the subspace, $\mathbf{Y} = \{y_1, y_2, \cdots, y_N\}^T$. L is the number of categories. λ_1 and λ_2 are weighting parameters. G_c is the c-th auxiliary projection target center matrix for samples of each category. The auxiliary projection target center is the symmetric target center of the projection target center of other categories with respect to the projection target center of the current category. The selections of G_c are shown in Algorithm 1.

Algorithm 1 The selection of G_c.

Input: The target projection matrix of subspace: $\mathbf{Y} = \{y_1, y_2, \cdots, y_N\}^T$; The number of categories: L.
Output: The auxiliary projection target center matrix: G_c

```
A = Y
for all c = 1 to L − 1 do
  for all j = 1 to L do
    if c == 0 then
      B^0 = A^{j−1}
    else
      B^j = A^{j−1}
    end if
  end for
  A = B
  G_c = 2Y_j − A
end for
```

Note: B^j is the j-th column of B.

In Equation (14), the first term is used to learn the projection matrix, the second term is used for feature selection. The third item is used to expand the inter-class distance between different categories of samples and reduce the dimension of the projection target area.

According to the analysis of l_{21}-norm by He et al. [29], the second term is optimized. $\left\| \mathbf{U}_p \right\|_{21}$ is deduced to $\mathrm{Tr}\left(\mathbf{U}_p^T \mathbf{R}_p \mathbf{U}_p \right)$, where $\mathbf{R}_p = \mathrm{Diag}\left(\mathbf{r}_p \right)$, $\mathbf{r}_p$ is an auxiliary vector of l_{21}-norm. The i-th parameter of $\mathbf{r}_p$ is $r_p^i = \frac{1}{2\left\| \mathbf{u}_p^i \right\|_2}$, $\mathbf{u}_p^i$ is the i-th row vector of $\mathbf{U}_p$. To keep the denominator from being 0, an infinite decimal α is introduced, and α is not 0. r_p^i is rewritten as:

$$r_p^i = \frac{1}{2\sqrt{\left\| \mathbf{u}_p^i \right\|_2^2 + \alpha}} \tag{15}$$

Equation (14) is derived and the computational formula of the projection matrix is obtained as:

$$\mathbf{U}_p = \left(\mathbf{X}_p\mathbf{X}_p^T + \lambda_1\mathbf{R}_p + \lambda_2\mathbf{X}_p\mathbf{X}_p^T\right)^{-1}\left(\mathbf{X}_p\mathbf{Y} + \lambda_2\sum_{c=1}^{\mathrm{L}}\mathbf{X}_pG_c\right) \tag{16}$$

The projection matrix of different modalities is obtained through Equation (16), and the test sample of each modality are sent into the common subspace to acquire fusion features $\mathbf{X}_p^T\mathbf{U}_p$. The fusion features are used to research human action recognition.

4. Experiment

In this paper, the experiment runs on a desktop.The hardware is as follows: the main board is X370 Taichi, the CPU is 3.4 GHz R7 1700X, the graphics card is GTX 1660, and the memory is 16 GB. The software is Python 3.7 and Matlab 2018b. For each sequence, the average computing time of MCSTV is about 0.046 s, MSTM is about 1.239 s, and MSTM-HOG is about 0.043 s.

The experiments of proposed method were tested on Microsoft Research Action3D (MSR-Action3D) [5] and University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD) [30]. Support vector machine (SVM) is used as a classifier to classify samples and obtain the final recognition accuracy. The dimension of skeleton feature MCSTV is related to the number of frames N, so we use Fisher Vector (FV) [31] to normalize MCSTV to make it linearly separable. After processing, the size of MCSTV is changed into $2pK \times 1$. Where p is the number of rows of MCSTV, K is the clustering center of FV. In this paper, $p = 3$, $K = 128$, the size of MCSTV is 768×1.

4.1. Datasets and Experimental Settings

4.1.1. MSR-Action3D

There are 557 depth sequences in this MSR-Action3D. The dataset includes 20 actions performed by 10 subjects, and each subject performs each action 2 to 3 times. The 20 actions are high wave, horizontal wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pick up throw. The positive order actions of the dataset are denoted as D1. The positive order actions and reverse order actions of the dataset are denoted as D2. In this paper, two different settings are employed to this dataset.

Setting 1. Similar to Li et al. [5], D1 and D2 are divided into 3 groups (AS1, AS2, AS3). In addition, the MSR-Action3D subset dataset is shown in Table 1. The actions with high similarity are divided into the same group, in order to evaluate the model performance traits according to the training dataset size changes, each group of samples is tested to three experiments. In Test1, 1/3 samples are used as training data, and the remaining samples are used as test data. In Test2, 2/3 samples are used as training data, and the remaining samples are used as test data. In Test3, half subjects are used for training and the rest ones used for testing.

Table 1. MSR-Action3D subset dataset

	Action Set 1 (AS1)	**Action Set 2 (AS2)**	**Action Set 3 (AS3)**
Action name	horizontal wave	high 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
	pick up throw	side boxing	pick up throw

Setting 2. Similar to Chen et al. [20], all samples in the MSR-Action3D are classified at the same time. The samples of subject 1, 3, 5, 7, 9 are used as training data, and samples of subject 2, 4, 6, 8, 10 are used as test data.

4.1.2. UTD-MHAD

There are 861 depth sequences in UTD-MHAD. The dataset includes 27 actions performed by 8 subjects (4 females and 4 males). Each subject performs each action 4 times. The 27 actions are swipe left, swipe right, wave, clap, throw, arm cross, basketball shoot, draw x, draw circle clockwise, draw circle counter clockwise, draw triangle, bowling, boxing, baseball swing, tennis swing, arm curl, tennis serve, push, knock, catch, pickup throw, jog, walk, sit2stand, stand2sit, lunge, squat. The positive order actions of the dataset are denoted as D3. The positive order actions and reverse order actions of the dataset are denoted as D4. In this paper, two different settings are employed to this dataset.

Setting 3. In order to evaluate the model performance traits according to the training dataset size changes, the samples in D3 and D4 are tested to three experiments. In Test1, 1/3 samples are used as training data, and the remaining samples are used as test data. In Test2, 2/3 samples are used as training data, and the remaining samples are used as test data. In Test3, we divid the samples into 5 parts and take turns to use 4 parts for training and 1 part for testing, the final result is the average of the 5 recognition rates.

Setting 4. Similar to Xu et al. [30], all samples in the UTD-MHAD are classified at the same time. The samples of subject 1, 3, 5, 7 are used to train, and samples of subject 2, 4, 6, 8 are used to test.

4.2. Parameter Selection

4.2.1. Parameter Selection of ξ

In the calculation of MSTM, we must use key information extraction based on inter-frame energy fluctuation algorithm to extract the key information in the motion energy. The amount of key information directly affects the ability of MSTM. Therefore, we need to set the appropriate ξ, which cannot only remove redundant information, but also retain key information completely. When setting different ξ, the effect of key motion energy retained is shown in Figure 9.

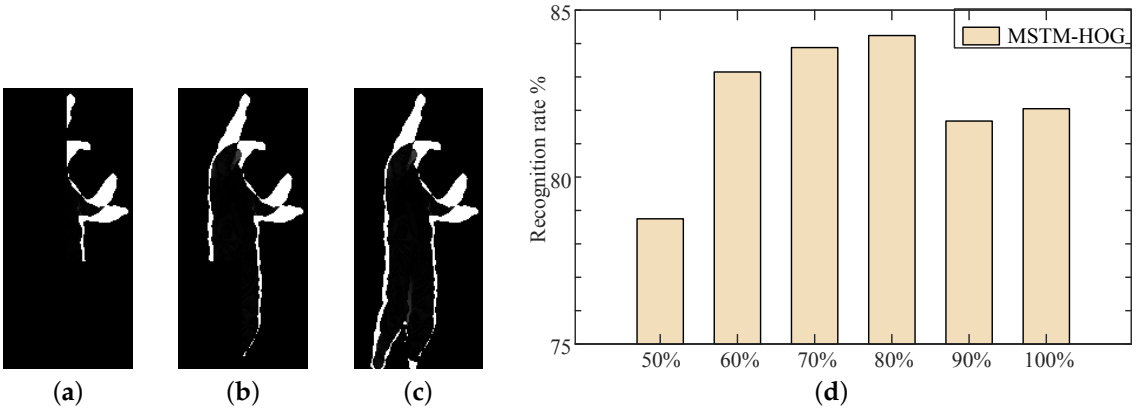

Figure 9. When setting different ξ, the effect of key motion energy retained. (**a**) $\xi = 50\%$; (**b**) $\xi = 80\%$ (**c**) The original motion energy (**d**) The recognition result of MTSM-HOG when setting different ξ.

Figure 9a shows the key motion energy of tennis serve retained when $\xi = 50\%$. It can be seen that the value of ξ is too small, much key information is mistaken for redundant, and the possibility of this action being recognized as tennis serve is reduced. Figure 9b shows the key motion energy of tennis serve retained when $\xi = 80\%$. Compared with the original motion energy in Figure 9c, it not only retains the key information, but also removes the energy of the area that the motion is not obvious. Figure 9d shows the recognition rate of MTSM-HOG when setting different ξ according to Setting 2.

It can be seen that MSTM-HOG achieves the highest recognition rate when $\xi = 80\%$. So we set ξ to 80% in the following experiment.

4.2.2. Selection of Image Features

Due to excessive noise, if the proposed MSTM is directly classified, the recognition results are affected. In this paper, we select Histogram of Oriented Gridient (HOG) [32] operator and Local Binary Pattern (LBP) [33] operator for feature extraction of image because they are not sensitive to light. HOG operator uses the image unit of 10×10 pixels to segment the image, combines every 2×2 image units into an image block, and slides the image block of 10 pixels to extract the HOG features of the image. LBP operator uses sampling radius of 2 and sampling point of 8 to extract the LBP features of the image. According to Setting 1, the results on D1 are shown in Figure 10 when HOG features and LBP features of MSTM are extracted.

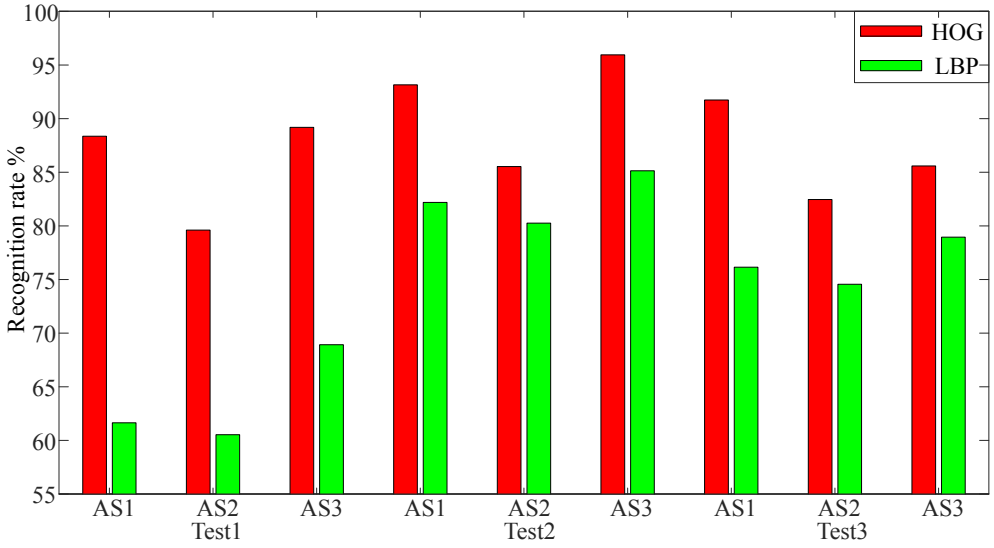

Figure 10. The recognition rate of MSTM when HOG features and LBP features are extracted.

As showed in Figure 10, the recognition accuracy of MSTM when extracting HOG features is higher than LBP. The LBP features mainly reflects the texture information around the pixel, while the HOG features can capture the image contour, and the main information of MSTM distributes in the image contour. Therefore, HOG features are more suitable for MSTM than LBP. In the following experiments, we only extract the HOG features of the images.

4.2.3. Parameter Selection of λ_1 and λ_2

When using MTSL to fuse image and skeleton features, we should select the fitting parameters λ_1 and λ_2. Assuming that $\lambda_2 = 0.01$, the optimal λ_1 is selected by enumerating different λ_1 and taking the recognition accuracy of our method as the evaluation standard. As can be seen from Figure 11a, when $\lambda_1 = 15$, our method achieves the optimal result. To obtain the optimal value of λ_2, we select the optimal value by enumerating different λ_2 when $\lambda_1 = 15$, and taking the recognition accuracy of our method as the evaluation criterion. As can be seen from Figure 11b, when $\lambda_2 = 0.05$, our method achieves the optimal result. In this experiment, the results are the recognition rate of our method on MSR-Action3D according to Setting 2.

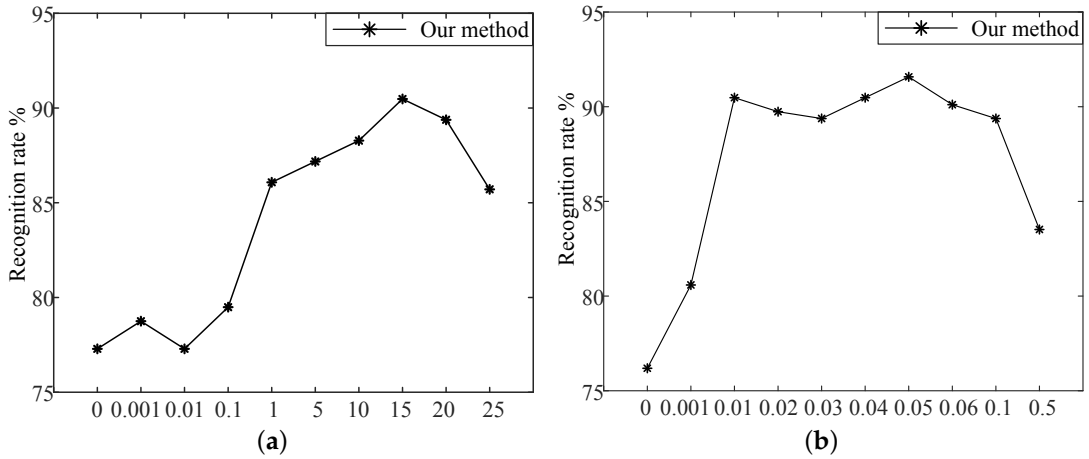

Figure 11. Parameter selection of λ_1 and λ_2. (**a**) Parameter selection of λ_1; (**b**) Parameter selection of λ_2.

4.3. Results and Discussion

4.3.1. Evaluation of MCSTV

MCSTV is a comprehensive vector formed by weighted accumulation of limb motion vectors. To verify that the accumulation of motion vectors can improve the action representation, we compare the recognition rate of each limb's movement features with MCSTV. According to Setting 2, the results of different movement features on MSR-Action3D are shown in Figure 12a. According to Setting 4, the results on UTD-MHAD are shown in Figure 12b.

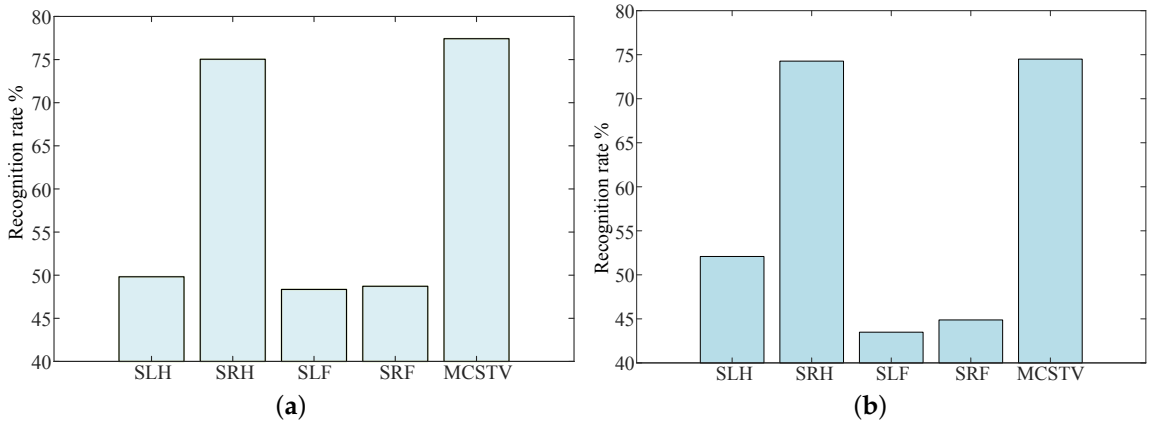

Figure 12. The recognition rate of different movement features. (**a**) The results on MSR-Action3D; (**b**) The results on UTD-MHAD.

As shown in Figure 12a,b, MCSTV achieves the highest recognition rate on two datasets. Among them, it reaches 77.42% on MSR-Action3D and 74.5% on UTD-MHAD, both of them are higher than other movement features.

Next, we compare the variations of the three axes of various movement features. We select the actions with the right hand as the main movement limb and compares the expressive effect of SRH, MCSTV_1 and MCSTV_2. Where MCSTV_1 is formed by the direct accumulation of each limb's motion vectors, MCSTV_2 is formed by the weighted accumulation of each limb's motion vectors. We select high wave and tennis serve in MSR-Action3D, and select baseball swing in UTD-MHAD. Various movement features of each action are shown in Figure 13.

The main motion limb of the high wave is right upper limb, so the trajectory of MCSTV should be similar to SRH. Tennis serve and baseball swing are the actions that the movement of each limb is obvious, and right upper limb is the main motion joint, so the final MCSTV trajectory should be dominated by SRH. As shown in Figure 13, compared with $MCSTV_1$, the trajectory of $MCSTV_2$ is closer to SRH. In particular, the trajectory of baseball swing's $MCSTV_2$ is similar to SRH, but $MCSTV_1$ is different. It can be explained that MCSTV formed by weighted accumulation is more accurate and highlight the main moving limbs.

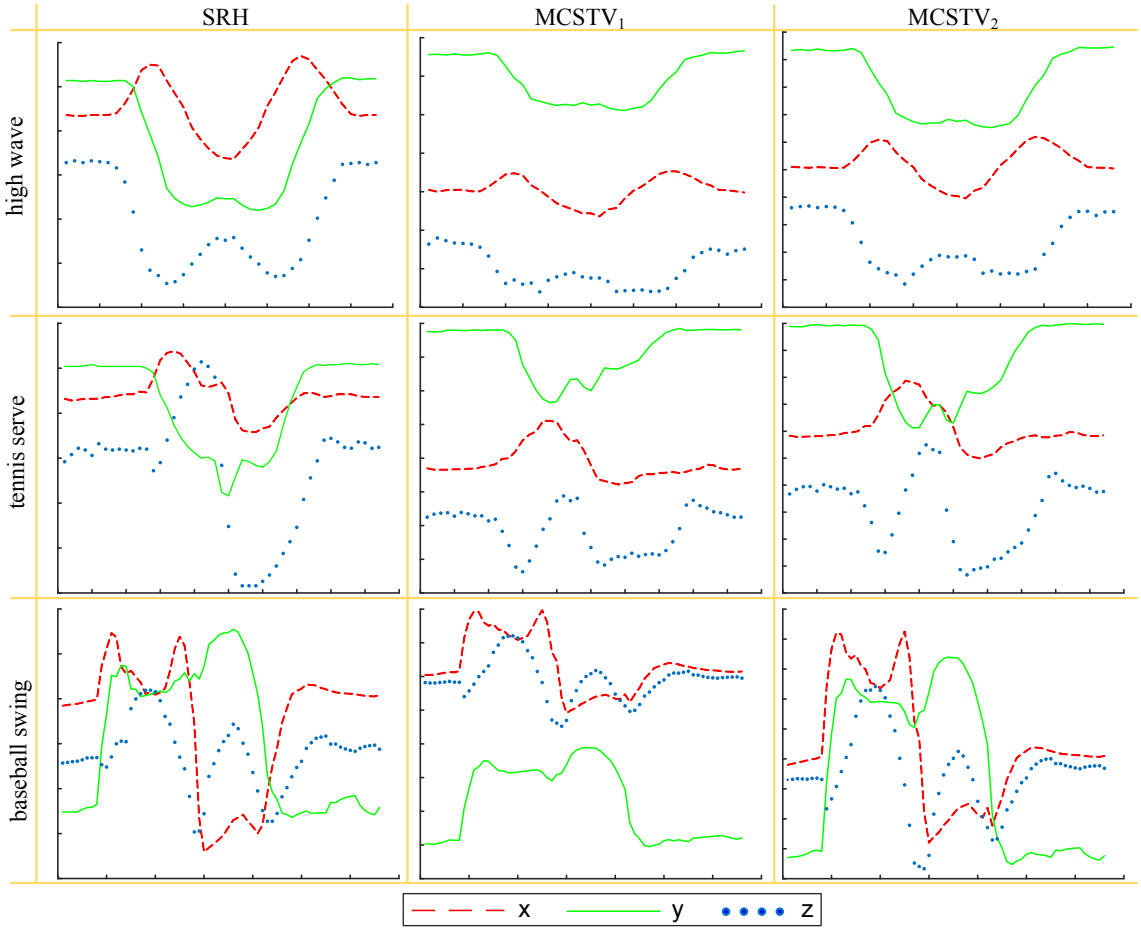

Figure 13. Various movement features of each action.

Then, we verify that MCSTV formed by the weighted accumulation describes actions more accurately. The recognition rates are used as the criterion, $MCSTV_2$ is compared with SRH and $MCSTV_1$. The results on MSR-Action3D are shown in Figure 14a, and the results on UTD-MHAD are shown in Figure 14b.

The data in Figure 14a,b are the recognition rates of various features running 20 times. The figure clearly shows the mean, maximum, minimum, median, outlier, upper quartile, and lower quartile of the multiple results. It can be seen from the two figures that the recognition rate of $MCSTV_2$ is higher than $MCSTV_1$ and slightly higher than SRH. The main reason is that MCSTV formed by weighted accumulation not only considers the motion vector of each limb, but also gives different weight to each limb according to its contribution, which can highlight the information of the main moving limbs and describe actions more accurately.

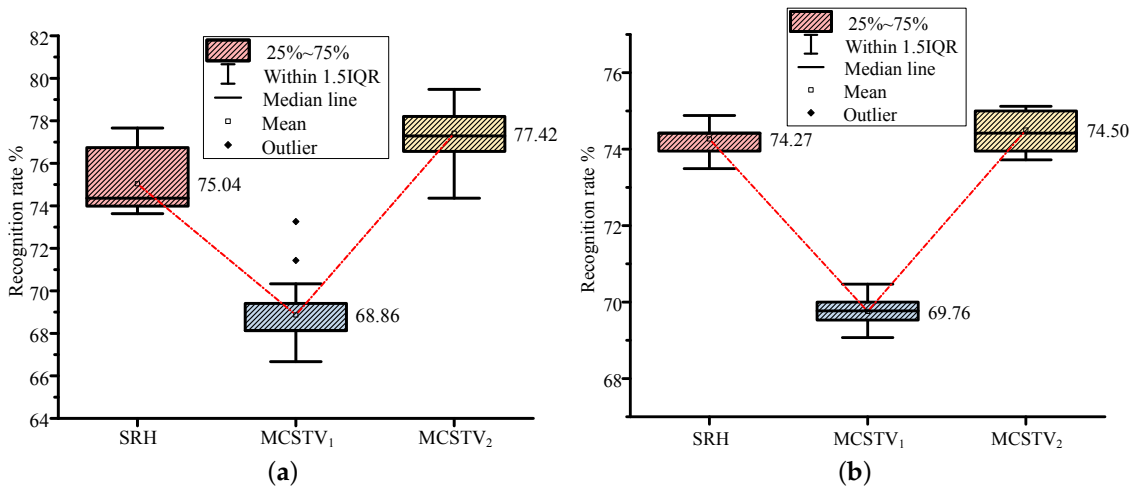

Figure 14. The recognition rate of SRH, $MCSTV_1$ and $MCSTV_2$. (**a**) The results on MSR-Action3D; (**b**) The results on UTD-MHAD.

4.3.2. Evaluation of MSTM

MSTM expresses the change of motion energy with time on three orthogonal axes. It retains spatial structure and temporal information of actions. To verify that MSTM can completely retain the spatial structure of actions, the recognition rate of MSTM is compared with MEI, MHI and DMM when only positive order actions exist. In this experiment, according to Setting 1, the results of different methods on D1 are shown in Table 2. According to Setting 3, the results of different methods on D3 are shown in Table 3.

Table 2. The recognition rates (%) of different methods on D1.

Method	Test1				Test2				Test3			
	AS1	AS2	AS3	Average	AS1	AS2	AS3	Average	AS1	AS2	AS3	Average
MEI-HOG	73.29	73.03	72.30	72.87	86.30	86.84	90.54	87.89	86.24	81.58	71.17	79.66
MHI-HOG	69.86	64.47	72.97	69.10	86.30	88.16	90.54	88.33	83.49	81.58	72.07	79.05
DMM-HOG	76.03	71.71	77.70	75.15	87.67	86.84	94.59	89.10	84.40	85.09	75.68	81.72
MSTM-HOG	88.36	79.61	89.19	85.72	93.15	85.53	95.95	91.54	91.74	82.46	85.59	86.60

Table 3. The recognition rates (%) of different methods on D3.

Method	Test1	Test2	Test3
MEI-HOG	69.51	83.28	89.53
MHI-HOG	72.47	89.55	94.19
DMM-HOG	78.57	94.08	98.84
MSTM-HOG	85.89	94.77	96.51

It can be seen from Tables 2 and 3 that MSTM-HOG achieves or approaches the highest recognition accuracy in most tests. The reason is that MSTM represents the projection of motion energy on three orthogonal axes, which completely preserves the spatial structure of actions. By contrast, the recognition accuracy of MEI and MHI was lower in multiple tests. Due to MEI and MHI describe the largest contour of actions, there is coverage of front and behind motion information, MEI and MHI lose the motion information inside the contour. In addition, MEI and MHI do not involve the depth information of actions. DMM achieves the highest recognition accuracy in some tests, mainly it accumulates the motion energy from three views and completely retains the spatial information.

Next, we verify that MSTM has a strong ability to represent temporal information, the recognition rate of MSTM is compared with MEI, MHI and DMM when positive and reverse order actions exist. According to Setting 1, the results of different methods on D2 are shown in Table 4. According to Setting 3, the results of different methods on D4 are shown in Table 5.

Table 4. The recognition rates (%) of different methods on D2.

Method	Test1				Test2				Test3			
	AS1	AS2	AS3	Average	AS1	AS2	AS3	Average	AS1	AS2	AS3	Average
MEI-HOG	29.11	33.88	29.05	30.68	30.14	25.66	33.78	29.86	36.99	35.53	32.43	34.98
MHI-HOG	37.67	39.80	36.49	37.97	40.41	48.68	45.27	44.79	44.75	48.25	38.29	43.76
DMM-HOG	34.93	34.87	27.70	32.50	32.19	32.24	31.08	31.84	38.36	35.52	30.63	34.84
MSTM-HOG	80.48	69.74	84.46	78.22	83.56	83.55	91.89	86.33	85.84	81.58	90.54	85.99

Table 5. The recognition rates (%) of different methods on D4.

Method	Test1	Test2	Test3
MEI-HOG	28.66	33.97	32.85
MHI-HOG	44.60	60.63	66.86
DMM-HOG	38.85	39.72	35.47
MSTM-HOG	83.80	92.68	94.48

In this paper, positive order high throw and reverse order high throw are considered to be two different actions. They have the same spatial trajectory and the opposite temporal information. So the number of actions in D2 is twice in D1, the number of actions in D4 is twice in D3. It can be known from Tables 4 and 5 that the recognition rate of each method is lower than that of Tables 2 and 3. MSTM maintains the highest recognition rate in all tests. The main reason is that MSTM splices the motion energy according to the temporal series, which can fully express the temporal information of actions. The MSTM of positive and reverse order actions are symmetric along the time axis, so the two actions can be accurately classified. By contrast, the recognition rate of MEI and DMM is lower in each test. The reason is that MEI and DMM could not express the temporal information. MEI and DMM of positive and reverse order actions are very similar and cannot be distinguished. MHI expresses part of the temporal information through brightness attenuation, so the recognition rate of MHI is higher than MEI and DMM, but far lower than MSTM.

4.3.3. Evaluation of Feature Fusion

MCSTV can accurately describe the integrity and cooperativity of human limbs. MSTM can completely record the spatial structure and temporal information of actions. To combine the advantages of MSTM and MCSTV, we use MTSL to fuse MCSTV and MSTM-HOG. To prove that the fusion feature describes actions more accurately, we compare the recognition accuracy of the fusion algorithm with single algorithms. According to Setting 2, the results on MSR-Action3D are shown in Table 6. According to Setting 4, the results on UTD-MHAD are shown in Table 7.

It can be seen from Tables 6 and 7 that the recognition accuracy of feature fused by MTSL algorithm is higher than single algorithms. The reason is that MTSL projects different features into a common subspace and complements the advantages of each single features.

It can be also known that the recognition accuracy of feature fused by MTSL is higher than CCA and DCA. Mainly, the MTSL algorithm constructs multiple projection targets to make the subspace samples converge to the hyperplane near the multiple projection target centers and increases the distance between the subspace samples. However, CCA and DCA mainly describe the correlation of two features, image and skeleton are two different modals with small correlation.

Table 6. The results on MSR-Action3D, according to Setting 2.

Method	Recognition Rate (%)
MEI-HOG	69.23
MHI-HOG	69.96
DMM-HOG	85.35
MSTM-HOG	84.24
MCSTV	77.42
CCA + MCSTV + MSTM-HOG	67.77
DCA + MCSTV + MSTM-HOG	75.09
Our Method	91.58

Table 7. The results on UTD-MHAD, according to Setting 2.

Method	Recognition Rate (%)
MEI-HOG	49.77
MHI-HOG	57.21
DMM-HOG	79.07
MSTM-HOG	83.72
MCSTV	74.50
CCA + MCSTV + MSTM-HOG	63.95
DCA + MCSTV + MSTM-HOG	77.21
Our Method	89.53

4.4. Comparison with Other Methods

We compare our method with other methods. According to Setting 2, the recognition accuracy comparison with other methods on MSR-Action3D is shown in Table 8, According to Setting 4, the recognition accuracy comparison with other methods on UTD-MHAD is shown in Table 9.

Table 8. The recognition accuracy comparison with other methods on MSR-Action3D.

Method	Recognition Rate (%)
DMM-HOG [7]	85.5
HON4D [6]	88.89
Bag of 3D points [5]	74.7
DMM-GLAC-FF [21]	89.38
Random Occupancy Patterns [23]	86.5
Depth Cuboid [24]	89.3
STOP [25]	84.8
Our Method	91.58

Table 9. The recognition accuracy comparison with other methods on UTD-MHAD.

Method	Recognition Rate (%)
Kinect [30]	66.1
Kinect + Inertial [30]	79.1
3DHoT-MBC [22]	84.4
Our Method	89.53

It can be seen from Tables 8 and 9 that the recognition accuracy of our method reached 91.58% on the MSR-Action3D and 89.53% on the UTD-MHAD, both of which were higher than the recognition accuracy of other methods listed. The evaluation results indicate the superiority of our method.

5. Conclusions

In this paper, we propose an action feature representation that considers the integrally cooperative movement features of human action, called MCSTV and MSTM. MCSTV accumulates weighted limbs' motion vector to form a new vector and uses this vector to account for the movement features of actions. MSTM algorithm projects key motion energy that extracted by key information extraction based on inter-frame energy fluctuation to three orthogonal axes and stitches them in temporal series to construct the MSTM. To describe the action information more accurately, the MTSL algorithm is used to fuse MCSTV and MSTM-HOG. The experimental results on MSR-Action3D and UTD-MHAD show that MCSTV not only considers the integral and cooperative between the motion joints, but also highlights the main moving limbs of the body. Compared with MEI, MHI, and DMM, MSTM describes the spatial structure and temporal information better. The recognition accuracy of features fused by MTSL algorithm is higher than most existing algorithms.

6. Future Expectations

When we use key information extraction based on inter-frame energy fluctuation algorithm to extract the key information, in some cases, the redundancy cannot be effectively removed because the habitual shaking of some joints is too sharp. Next, we will focus on how to effectively remove the redundant information.

Author Contributions: Conceptualization, X.C., Z.H., J.L. and T.Y.; methodology, X.C. and Z.H.; validation, X.C.; writing—original draft preparation, X.C.; writing—review and editing, X.C., Z.H., J.L. and T.Y.; funding acquisition, Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (61063021), Project of scientific research innovation plan for Postgraduates in Jiangsu Province (KYCX20-2551) and Zhejiang Public Welfare Technology Research Social Development Project (2017C33223).

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

References

1. Ji, S.W.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
2. Xu, Y.; Hou, Z.J.; Liang, J.Z.; Chen, C.; Jia, L.; Song, L. Action recognition using weighted fusion of depth images and skeleton's key frames. *Multimed. Tools Appl.* **2019**, *78*, 25063–25078. [CrossRef]
3. Chen, C.; Jafari, R.; Kehtarnavaz, N. Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 51–61. [CrossRef]
4. Arshad, H.; Khan, M.A.; Sharif, M.; Yasmin, M.; Javed, M.Y. Multi-level features fusion and selection for human gait recognition: An optimized framework of Bayesian model and binomial distribution. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 3601–3618. [CrossRef]
5. Li, W.Q.; Zhang, Z.Y.; Liu, Z.C. Action recognition based on a bag of 3D points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA, 13–18 June 2010.
6. Oreifej, O.; Liu, Z.C. Histogram of oriented 4D normals for activity recognition from depth sequences. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
7. Yang, X.D.; Zhang, C.Y.; Tian, Y.L. Recognizing actions using depth motion maps-based histograms of oriented gradients. In Proceedings of the 20th ACM International Conference on Multimedia, Machinery, New York, NY, USA, 29 October–2 November 2012; pp. 1057–1060.
8. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.

9. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3D skeletons as points in a lie group. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.
10. Luvizon, D.C.; Tabia, H.; Picard, D. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognit. Lett.* **2017**, *99*, 13–20. [CrossRef]
11. Hardoon, D.R.; Szedmak, S.; Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **2004**, *16*, 2639–2664. [CrossRef] [PubMed]
12. Haghighat, M.; Abdel-Mottaleb, M.; Alhalabi, W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1984–1996. [CrossRef]
13. Xi, X.; Tang, M.; Miran, S.M.; Luo, Z. Evaluation of Feature Extraction and Recognition for Activity Monitoring and Fall Detection Based on Wearable sEMG Sensors. *Sensors* **2017**, *17*, 1229. [CrossRef] [PubMed]
14. Cao, X.H.; Xiong, T.; Jiao, L.C. Supervised Band Selection Using Local Spatial Information for Hyperspectral Image. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 329–333. [CrossRef]
15. Hossain, M.R.I.; Little, J.J. Exploiting Temporal Information for 3D Human Pose Estimation. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
16. Lv, F.J.; Nevatia, R. Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 359–372.
17. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
18. 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 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2752–2759.
19. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
20. Chen, C.; Jafari, R.; Kehtarnavaz, N. Action recognition from depth sequences using depth motion maps-based local binary patterns. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015.
21. Chen, C.; Hou, Z.J.; Zhang, B.C.; Jiang, J.J.; Yang, Y. Gradient local auto-correlations and extreme learning machine for depth-based activity recognition. *Int. Symp. Vis. Comput.* **2015**, *9474*, 613–623.
22. Zhang, B.C.; Yang, Y.; Chen, C.; Yang, L.L.; Han, J.G.; Shao, L. Action recognition using 3d histograms of texture and a multi-class boosting classifier. *IEEE Trans. Image Process.* **2017**, *26*, 4648–4660. [CrossRef] [PubMed]
23. Wang, J.; Liu, Z.C.; Chorowski, J.; Chen, Z.Y.; Wu, Y. Robust 3D Action Recognition with Random Occupancy Patterns. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012.
24. Xia, L.; Aggarwal, J.K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
25. Vieira, A.W.; Nascimento, E.R.; Oliveira, G.L.; Liu, Z.C.; Campos, M.F.M. STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences. In Proceedings of the CIARP 2012: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Buenos Aires, Argentina, 3–6 September 2012.
26. Zhang, Z.; Zhao, M.B.; Chow, T.W.S. Binary and multi-class group sparse canonical correlation analysis for feature extraction and classification. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 2192–2205. [CrossRef]
27. Wang, K.Y.; He, R.; Wang, L.; Wang, W.; Tan, T.N. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2010–2023. [CrossRef] [PubMed]
28. Eswaraiah, R.; Reddy, E.S. Robust medical image watermarking technique for accurate detection of tampers inside region of interest and recovering original region of interest. *IET Image Process.* **2015**, *9*, 615–625. [CrossRef]

29. He, R.; Tan, T.N.; Wang, L.; Zheng, W.S. ℓ_{21} Regularized correntropy for robust feature selection. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
30. Chen, C.; Jafari, R.; Kehtarnavaz, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015.
31. Luo, G.; Wei, J.T.; Hu, W.M.; Maybank, S.J. Tangent Fisher Vector on Matrix Manifolds for Action Recognition. *IEEE Trans. Image Process.* **2019**, *29*, 3052–3064. [CrossRef] [PubMed]
32. Fan X.J.; Tjahjadi, T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognit.* **2015**, *48*, 3407–3416. [CrossRef]
33. Zhang, M.; Muramatsu, C.; Zhou, X.R.; Hara, T.; Fujita, H. Blind Image Quality Assessment Using the Joint Statistics of Generalized Local Binary Pattern. *IEEE Signal Process. Lett.* **2015**, *22*, 207–210. [CrossRef]