Computer Science and Information Systems 19(3):1445–1462 https://doi.org/10.2298/CSIS210322042W

Human Action Recognition Using a Depth Sequence Key-frames Based on Discriminative Collaborative Representation Classifier for Healthcare Analytics

Yuhang Wang[1], Tao Feng[2,3], and Yi Zheng[1,⋆]

[1] Institute of physical culture, Harbin University
Harbin 150000, China
83008943@qq.com
86959936@qq.com
[2] Department of Physical Education, Harbin Finance University
Harbin,150000, China
ancrum@qq.com
[3] The Graduate School of Saint Paul University Philippines
Ottawa, Philippines

Abstract. Using deep map sequence to recognize human action is an important research field in computer vision. The traditional deep map-based methods have a lot of redundant information. Therefore, this paper proposes a new deep map sequence feature expression method based on discriminative collaborative representation classifier, which highlights the time sequence of human action features. In this paper, the energy field is established according to the shape and action characteristics of human body to obtain the energy information of human body. Then the energy information is projected onto three orthogonal axes to obtain deep spatial-temporal energy map. Meanwhile, in order to solve the problem of high misclassification probability of similar samples by collaborative representation classifier (CRC), a discriminative CRC (DCRC) is proposed. The classifier takes into account the influence of all training samples and each kind of samples on the collaborative representation coefficient, it obtains the highly discriminative collaborative representation coefficient, and improves the discriminability of similar samples. Experimental results on MSR_Action3D data set show that the redundancy of key-frame algorithm is reduced, and the operation efficiency of each algorithm is improved by 20%-30%. The proposed algorithm in this paper reduces the redundant information in deep map sequence and improves the extraction rate of feature map. It not only preserves the spatial information of human action through the energy field, but also records the temporal information of human action in a complete way. What's more, it still maintains a high recognition accuracy in the action data with temporal information.

Keywords: action recognition, deep map sequence, deep spatial-temporal energy map, discriminative CRC, energy information.

1. Introduction

Human action recognition is a research hotspot in machine vision and artificial intelligence [1,2]. Many research achievements have been applied in the fields of human-

[⋆] Corresponding author

computer interaction, biometrics, health monitoring, video surveillance system, motion sensing games, robotics, etc,. Most of the early researches on action recognition were carried out on color video sequences collected by RGB cameras. For example, motion energy image (MEI) and motion history image (MHI) proposed by Bobick and Davis [3] were originally color videos collected by RGB cameras. MEI represents the outline of human action and does not involve the time sequences information of human action. MHI expresses temporal information and spatial contour of human action through brightness attenuation. However, due to the action occlusion, some action information is missing, and the final expressed time sequence information is incomplete. Due to the existence of redundant frames, the gray value of the final generated MHI is relatively concentrated near the redundant frames, which affects the final recognition accuracy [4].

With the development of imaging technology, especially the introduction of depth camera, the research object of human action recognition is transformed from the original RGB image to the depth image. Compared with the previous RGB images, the depth map sequences collected by the structured light depth sensor are not sensitive to light changes, and provide depth human action data. So far, researchers have done a lot of researches on depth map sequences. Zhu et al. [5] proposed 3D points, using a small amount of 3D points to represent human action. Luo et al. [6] proposed the depth cube and established a novel depth cube similarity feature to describe the local 3D depth cube around the depth map sequence. Xuan et al. [7] proposed surface normals and used 4-dimensional surface normal-direction histogram descriptors to capture the structural information of human action changes. Nie et al. [8] proposed bone joints, which were used to represent human action. Chaudhary et al. [9] used a depth motion map (DMM) to represent human action. Where, DMM was to project the depth map onto three orthogonal Cartesian planes, generated 2D projection maps from three perspectives according to the front view, side view and top view, and accumulated the image difference between two continuous projection maps to generate DMM from three perspectives. Mattiev et al. [10] proposed new associative classifiers, called DC, DDC and CDC, that used distance-based agglomerative hierarchical clustering as a post-processing step to reduce the number of its rules, and in the rule-selection step, it used different strategies (based on database coverage and cluster center) for each algorithm. Human action is composed of spatial information and temporal information. Spatial information reflects the spatial distribution of human body information, and temporal information reflects the sequence of human body information. DMM completely describes the spatial information of human action, but it cannot describe the temporal information of depth map sequence. When there are some actions with the same space trajectory and opposite time sequence in the database, the generated feature map is the same, but the two actions cannot be distinguished.

Although human action recognition has made great progress in recent years, it still has many shortcomings. In this paper, a key frame algorithm is proposed to solve the problem of too much redundant information in depth map sequence. Firstly, the redundancy coefficient is utilized to describe the redundancy. Then, according to the sequence of redundant coefficients, the redundant frames in the depth map sequence are located and deleted to obtain the key frame sequence to express human action sufficiently. Our main contributions are as follows:

1. In this paper, a new depth spatial-temporal energy feature expression (abbreviated as DSTEFE) method is proposed to solve the problem of poor temporal information of feature maps extracted from depth map sequences.
2. The energy field is established according to the shape and action characteristics of human body to obtain the energy information of human body. Then the energy information is projected onto three orthogonal axes to obtain deep spatial-temporal energy map.
3. Meanwhile, in order to solve the problem of high misclassification probability of similar samples by collaborative representation classifier (CRC), a discriminative CRC (DCRC) is proposed. The classifier takes into account the influence of all training samples and each kind of samples on the collaborative representation coefficient, it obtains the highly discriminative collaborative representation coefficient, and improves the discriminability of similar samples. This proposed method highlights the action information of human body and further improves the accuracy of action recognition.

This paper is organized as follows. Section 2 introduces the related works. Section 3 detailed illustrate the proposed DSTEFE based on DCRC for human action recognition. Section 4 gives the experiments for the proposed method. Finally, a conclusion is conducted in Section 5.

2. Related Works

In the early stage of human action recognition, people usually use RGB camera to collect the color video sequence of human action, and then extract the feature map from the color video sequence. MEI is initially extracted from the color video sequence. Firstly, the foreground area of human action is extracted, and binarization is carried out to obtain the binarization image sequence $B(x,y,t)$. Then, the union set of the binary image sequence is evaluated to obtain the feature graph of MEI [11]. The calculation of MEI is as follows:

$$M_\delta(x,y,t) = \cup_{i=0}^{\delta-1} B(x,y,t-i). \tag{1}$$

Where $M_\delta(x,y,t)$ represents MEI generated by δ images at frame t in the video sequence. x and y represent the height and width values of one point in the image respectively. t denotes the serial number of a frame in the image sequence.

MEI expresses the spatial contour of human action through the union of the binary foreground region. However, the video sequence of human action expressed in this way has the following problems: 1) MEI represents the maximum contour boundary of human action. Due to the occlusion of action information from front to back, some action information will be lost during movement; 2) MEI cannot express the time sequence information of human action. When there are actions in the database with the same spatial trajectory and opposite time sequence, the generated feature maps are the same and cannot be distinguished.

In order to show the human action, MHI is a feature map that can express some temporal information of human action. Different from MEI, MHI is a gray image. The gray value at each point is a time-history function. MHI can be represented by a simple substitution and attenuation operator, calculated as:

$$H_\tau(x,y,t) = \begin{cases} \tau & B(x,y,t) = 1 \\ max(0, H_\tau(x,y,t-1)) & others \end{cases} \quad (2)$$

Where $H_\tau(x,y,t)$ is the MHI generated by τ images at frame t in the video sequence. τ is the initial brightness. $B(x,y,t)$ is the binary image sequence.

Compared with MEI, MHI is significantly improved. It not only retains the spatial outline of human action, but also shows the temporal information of human action by brightness attenuation. But there are also some shortcomings: 1) There are many redundant frames in the collected video sequence, so that the gray value distribution of the final generated MHI is concentrated near the redundant frames, which seriously affects the accuracy of recognition; 2) The front and back occlusion of action information makes some action information missing , which makes it impossible to accurately express human action.

With the introduction of depth camera, people also begin to use depth map sequence for human action recognition research. Compared with the previous color video sequences, depth map sequences are not sensitive to light changes, so it is more convenient to extract the foreground area of human action and provides the depth information of human action. Yang et al. [12] proposed DMM, which projected each frame of depth map sequence onto three orthogonal Cartesian planes, and generated 2D projection images from three perspectives according to the front view, side view and top view, respectively. They were represented by map_f, map_s and map_t. DMM is calculated as:

$$S_v = \sum_{i=2}^{F}((|map_v^i - map_v^{i-1}| > \varepsilon). \quad (3)$$

Where $v \in f, s, t$ represents the projection angle of view. f, s and t denote the front view, side view and top view. S_v is the DMM of projection angle of view v. map_v^i is the $i-th$ frame graph of projection angle of view v. ε is the difference threshold. F is the frame number of the depth map sequence. $|map_v^i - map_v^{i-1}|$ represents the difference image of two consecutive projected images.

Compared with MEI, DMM fully uses the depth information of depth map sequence, but the sequence DMM information is also unable to express the temporal information of human action, and does not have the ability to distinguish positive and negative sequence actions.

3. Proposed DSTEFE Based on DCRC for Human Action Recognition

3.1. Proposed action recognition framework

The human action recognition framework based on DSTEFE and DCRC algorithm is shown in figure 1. Firstly, the redundant frames in depth map sequence are eliminated by the redundancy coefficient of the difference image sequence, and the key frame sequence sufficiently expressing human action is obtained. Then the energy field is established according to the shape and action characteristics of human body to obtain the energy information of human body. Then the energy information of human body is projected onto three orthogonal axes to obtain DSTEFE. Finally, HOG (histogram of Oriented gradient) features are extracted from each DSTEM and sent to a new DCRC classifier for human action recognition.

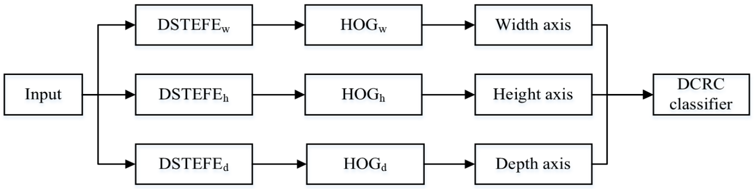


Fig. 1. Proposed human action recognition framework

3.2. Key frame algorithm

Due to the uneven human action rate during sampling, there are a large number of similar frames in the collected depth map sequences. In this paper, the similar frames appearing at the similar time in the depth map sequences are called redundant frames. After the redundant frames are eliminated, the remaining depth map sequences are called key frame sequences.

In the action recognition process, human action can be expressed only by the depth frame of the key position. However, there are a lot of redundant frames in the collected database, which has a great influence on the future research. Aiming at the above problems, the redundancy coefficient is proposed to describe the redundancy degree, and the key frame algorithm is further proposed based on the redundancy coefficient. By eliminating redundant frames of depth map sequence, redundant information is reduced, which makes the same action have approximate time interval, thus improving the operation rate of feature map and recognition accuracy.

The overall flow of the key frame algorithm is shown in figure 2.

1. The image difference between two adjacent frames of depth map sequences is obtained and the image difference sequence is generated.
2. The redundant frames in depth map sequence are located and deleted by the maximum redundancy coefficient.
3. Repeat the above steps until obtain the key frame sequence sufficiently expressing the human action.

This algorithm firstly executes difference processing between adjacent frames of depth map sequence, and then obtains the difference image of adjacent frames, which is calculated as:

$$D(x,y,t) = |I(x,y,t+1) - I(x,y,t)|. \tag{4}$$

Where $I(x,y,t)$ is the $i{-}th$ frame image of the original depth map sequence. $D(x,y,t)$ is the difference image between $(t+1)-th$ frame and $t-th$ frame of the original depth map sequence, namely the $t-th$ frame of the difference image sequence.

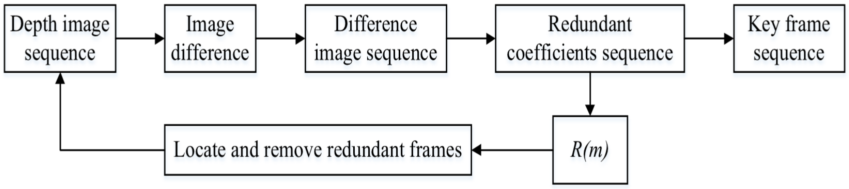

Depth image sequence
Image difference
Difference image sequence
Redundant coefficients sequence
Key frame sequence
R(m)
Locate and remove redundant frames

Fig. 2. Key frame flow chart

Then the redundancy coefficient of each frame in the difference image sequence is calculated to show the similarity between adjacent frames in the original depth map sequence. The calculation process of the redundancy coefficient of each frame in the differential image sequence is as follows.

First, the L2-norm of each frame in the difference sequence is calculated.

$$\alpha(t) = ||D(t)||_2 = \sqrt{\lambda_{max}(D^T(t)D(t))}. \tag{5}$$

Where $\alpha(t)$ represents the L2-norm of $t-th$ frame in the difference image sequence. λ_{max} is the maximum eigenvalue of a difference image.

Second, the L2-norm values of each frame in the difference image sequence are projected to the interval [0,1] to obtain the corresponding redundancy coefficient.

$$R(t) = e^{-\alpha(t)}. \tag{6}$$

Where $R(t)$ represents the redundancy coefficient of $t-th$ frame in the difference image sequence.

Third, the redundancy coefficients of each frame in the difference image sequence are sorted from big to small. It will find out the maximum redundancy coefficient $R(m)$ and its corresponding difference image frame $D(m)$. According to the difference image frame $D(m)$, it finds the corresponding redundant frames in the original depth map sequence and removes them. Repeat the above operations, remove redundant frames in the sequence, and obtain N frame sequences sufficiently expressing human action. N is determined by the experiment results.

3.3. Deep Spatial-temporal Energy Feature Expression

To solve the problem of missing time sequence information of generated feature map from depth map sequence, a feature expression method that can completely express action spatial-temporal sequence information is proposed, namely DSTEFE. DSTEFE reflects the change of human action energy information distribution on three orthogonal axes. Firstly, the energy field is established according to the shape and action characteristics of human body to obtain the energy information of human action. The energy information of human body is projected to 3 orthogonal Cartesian planes to generate 2d-projection images from 3 perspectives. Then two 2d-projection images are selected to continue to project onto the three orthogonal axes to generate a one dimension energy distribution list. DSTEFE with three orthogonal axes is formed after time order splicing. The three orthogonal axes are width axis (w), height axis (h) and depth axis (d), which correspond to the width direction, height direction and depth direction of the depth frame, respectively. L_w, L_h and L_d represent the corresponding one-dimension energy distribution list. The flow chart of DSTEFE is shown in figure 3.

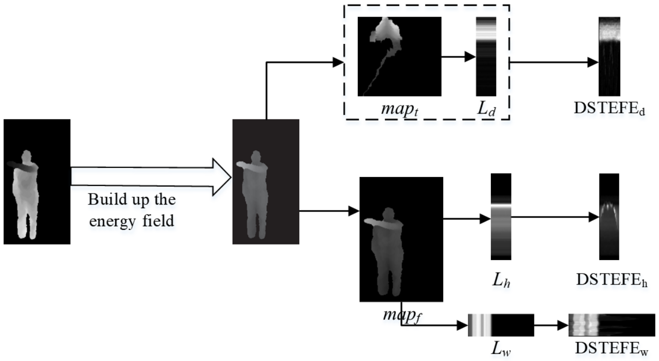


Fig. 3. The flow chart of DSTEFE

Step 1. Building up the energy field.

As shown in figure 4, the energy field of human body is first established to obtain the energy information of human action, so as to highlight the information of human action. The energy field coordinate system is shown in figure 4. It takes the height of the depth map as the x-axis direction and the width of the depth map as the y-axis direction.

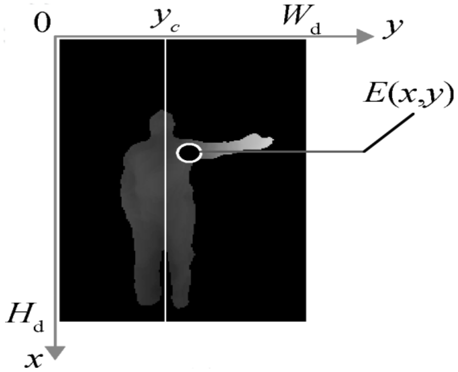


Fig. 4. Energy field coordinate system

According to the characteristics of forward stretching of human action, the depth distance between human foreground and the background is denoted as the forward energy of human body. It is calculated as:

$$E_f(x, y) = 255 - f(x, y). \tag{7}$$

Where $E_f(x, y)$ represents the forward energy of the human body. $f(x, y)$ is the depth value of the human body. According to the characteristics of lateral extension of human action, the distance between the foreground and the central axis of human body is denoted as the lateral energy of human body. It is calculated as:

$$E_s(x, y) = |y - y_c|. \tag{8}$$

Where $E_s(x, y)$ represents the lateral energy of the human body. When y_c is the initial frame of the action, the y-axis coordinate of the human body under the stand-at-attention posture is calculated as:

$$\sum_{i=0}^{H_d} \sum_{j=0}^{y_c} f(i, j) = 0.5 \sum_{i=0}^{H_d} \sum_{j=0}^{W_d} f(i, j). \tag{9}$$

Where W_d is the width of the depth map. H_d is the height of the depth map. Because there are many overlap areas between stretch up and foreground in the human action, this paper does not record the height direction energy of human body. The total energy E(x,y) of human body is calculated as:

$$E(x, y) = \sqrt{E_f^2(x, y) + E_s^2(x, y)}. \tag{10}$$

Because the forward energy and the lateral energy are linear operators, but the total energy is not linear operators. The absolute value is used to calculate the total energy.

$$E(x, y) = |E_f(x, y)| + |E_s(x, y)|. \tag{11}$$

The depth frame comparison of the energy field is shown in figure 5. Figure 5(a) shows the depth frame without the establishment of the energy field. Figure 5(b) shows the depth frame with the establishment of the energy field. Compared to figure 5(a), the establishment of energy field in figure 5(b) can significantly highlight the information of human action, which is conducive to enhancing the effect of human action recognition.

Step 2. Calculating DSTEFE.

The energy information of human body is projected to three Cartesian planes, and the 2D-energy projection diagram of three perspectives is generated according to the front view, side view and top view, which are represented by map_f, map_s and map_t respectively. In order to obtain the energy distribution of the width axis, height axis and depth axis in the action space, the front view and the top view are selected to continue to project onto the corresponding orthogonal axis. That is, the row sum or column sum of the two dimension energy projection graph is computed. Three one-dimension energy distribution lists are generated according to the width axis, height axis and depth axis, denoted as L_w, L_h and L_d, respectively. The formula is:

$$L_u(k) = \sum_{x=1}^{W_m} map_v(x, k) || \sum_{y=1}^{H_m} map_v(k, y). \tag{12}$$

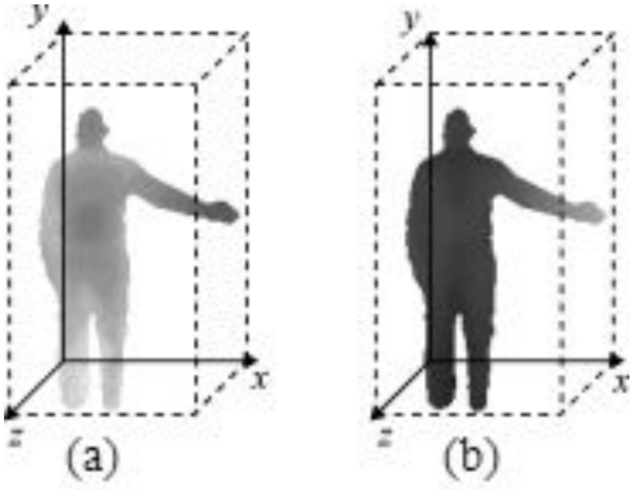


Fig. 5. The depth frame comparison

Where $v \in f, s, t$, $u = w, h, d$. w is width axis. h is height axis. d is depth axis. W_m is the width of the 2-dimension energy projection graph. H_m is the height of the 2-dimension energy projection graph. $L_u(k)$ is the k-th element of the projection list on the u axis.

$L_u(k)$ is normalized and spliced into DSTEFE of each axis in time order. For depth map sequence with N frames, DSTEFE is calculated as:

$$T_u(t) = L_u^t. \tag{13}$$

Where L_u^t represents the one-dimension energy distribution list of $t-th$ frame on the u-axis. T_u stands for DSTEFE on the u-axis. $T_u(t)$ denotes the $t-th$ row of T_u. Region of Interest (ROI) processing is carried out for each DSTEFE according to the maximum and minimum values of the width, height and depth in human action. That is, the image is cropped and the size is normalized.

3.4. Discriminative Collaborative Representation Classifier (DCRC)

Collaborative Representation Classifier (CRC) [13] is a very effective classifier, which is believed that the test samples can be approximately linearly represented by all training samples. Given a training sample set $D = D_1, \cdots, D_k, \cdots, D_K$ with K classes, where $D_k(k = 1, 2, \cdots, K)$ is the sample vector set corresponding to category K. y is used to represent a test sample, D collaboration can be represented as $y = D\alpha$, where α is the collaboration representation coefficient vector of the test sample. In this subsection, we introduce a new CRC classifier.

Assuming S is used to represent the linear space spanned by the collaboration of all the training samples. S_i represents the linear subspace spanned by a sample $D_i(i = 1, 2, \cdots, K)$ of the same class. $L = 1, 2, \cdots, K$ denotes the set of all categories.The test samples that do not belong to the space S can be represented as $y \approx D\alpha$, which can only indicate that the one category in test samples belongs to the one category in L. Then the residuals between the test samples and each category are reconstructed to approximate the category of the samples [14]. However, when the two classes in the training sample are very similar (such as D_i and D_j), the samples reconstructed by the corresponding coefficients α_i and α_j in the representational coefficient vectors obtained by CRC have a high similarity degree, so the probability of misclassification based on the residual classification rule will be increased.

In order to improve the discriminability of CRC for similar actions and improve the performance of the classifier, a highly discriminative cooperative representation coefficient is obtained by quadratic constraint for the coefficients, a DCRC classifier is proposed. First,a shared sample point $\hat{y} = D\alpha^* = D(\alpha_1^*, \cdots, \alpha_K^*)$ in space S is determined, where α^* is the corresponding representation coefficient vector of the sample point. The shared sample point should satisfy two conditions: 1) the similarity between the sample point and the test sample is very high; 2) The distance sum from the reconstructed sample point $\hat{y}_i = D_i\alpha_i^*$ to the sample point in each subspace S_i is the minimum. Then, after continuous optimization for the objective function, the optimal collaboration representation coefficient α^* and the optimal shared sample point can be obtained. Finally, it will be stopped until the residual difference between the shared sample point $\hat{y}$ and the reconstructed sample point $\hat{y}_i$ in a subspace is the smallest. So the category of test sample is obtained as shown in figure 6.

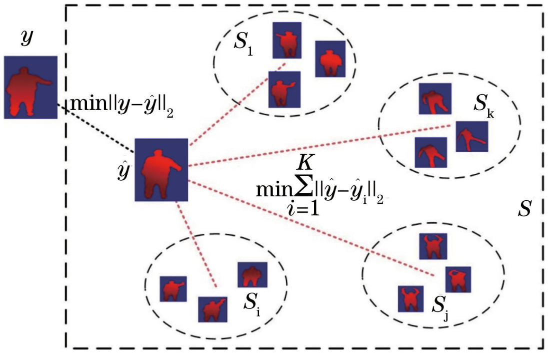


Fig. 6. DCRC process

It can be seen that the closer two samples denotes the greater probability that the two samples belong to a same category. Assuming that all samples are independently distributed in the space. $l(y)$ is used to represent the label of sample y, the probability of test sample y belonging to category i is:

$$\begin{aligned} P[l(y)=i] &= P[l(y)=l(\hat{y})|l(\hat{y})=i]P[l(\hat{y})=i] \\ &= P[l(y)=l(\hat{y})|l(\hat{y})=i]P[l(\hat{y})=i|l(\hat{y})\in L]P[l(\hat{y})\in L] \end{aligned} \tag{14}$$

Because the samples are independent of each other. So $P[l(y) = l(\hat{y})|l(\hat{y}) = i]P[l(\hat{y}) = l(y)|l(\hat{y}) \in L]$, formula (14) is equivalent to:

$$\begin{aligned} P[l(y) = i] &= P[l(y) = l(\hat{y})|l(\hat{y}) = i]P[l(\hat{y}) = i] \\ &= P[l(y) = l(\hat{y})]P[l(\hat{y}) = i|l(\hat{y}) \in L] \end{aligned} \tag{15}$$

Here, $P[l(y) = l(\hat{y})]$ can measure the distance between test sample y and $\hat{y}$, that is, $||y - \hat{y}||^2$. Because $\hat{y}_i$ falls inside the subspace S_i, so $P[l(\hat{y}) = i|l(\hat{y}) \in L]$ can be considered as measuring the distance between $\hat{y}$ and $\hat{y}_i$, namely, $\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||_2^2$. To obtain the label of the test sample, let

$$maxP[l(y) = i] = min(||y - D\alpha||_2^2 + \mu\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||_2^2). \tag{16}$$

In order to reduce the risk of over-fitting and computational complexity, Tikhonov matrix regularization term is used to constrain this function, and the final objective function is obtained:

$$\hat{\alpha} = argmin_{\alpha}||y - D\alpha||_2^2 + \lambda||\mathit{\Gamma}\alpha||_2^2 + \mu\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||_2^2. \tag{17}$$

Where λ and μ are the regularization constraint parameters. $\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||$ makes double constraint for α_i based on D_i, which can enhance the discriminability of the final coefficient vector α. when $\mu = 0$, the model of formula (17) becomes CRC. So μ must be larger than zero. Equations (14)-(17) prove the feasibility of DCRC from the perspective of probability. The category of test sample y can be determined by taking the maximum probability belonging to a single category.

The partial derivative of constraint $\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||$ for coefficient vector α is solved as follows:

$$\begin{aligned} \frac{\partial}{\partial\alpha}(\sum_{i=1}^{K} ||\hat{y} - D_i\alpha_i||_2^2) &= \frac{\partial}{\partial\alpha}\sum_{i=1}^{K} tr[(\hat{y} - D_i\alpha_i)^T(\hat{y} - D_i\alpha_i)] \\ &= \sum_{i=1}^{K} \frac{\partial}{\partial\alpha} tr(\alpha^T D^T D\alpha - \alpha^T D^T D_i\alpha_i - \alpha_i^T D_i^T D\alpha + \alpha_i^T D_i^T D_i\alpha_i) \end{aligned} \tag{18}$$

Let $\bar{D}_i = [0, \cdots, D_i, \cdots, 0]$, so formula (18) can be simplified as:

$$\sum_{i=1}^{K} 2D^T D\alpha - 2(D^T \bar{D}_i\alpha + \bar{D}_i^T D\alpha) + 2\bar{D}_i^T \bar{D}_i\alpha. \tag{19}$$

Combined with the optimal solution of CRC model, the optimal solution of the discriminant cooperative representation classifier can be obtained:

$$\hat{\alpha} = [D^T D + \lambda\mathit{\Gamma}^T\mathit{\Gamma} + \mu\sum_{i=1}^{K}(D - \bar{D})^T(D - \bar{D}_i)]^{-1}D^T y. \tag{20}$$

Finally, a new rule is used to determine the category of test sample:

$$e_i = ||D\hat{\alpha} - D_i\hat{\alpha}_i||_2^2. \tag{21}$$

$$label(y) = argmin_i[e(i)]. \tag{22}$$

4. Experiments and Analysis

This experiment is conducted on MATALB2017a, Python3.5, CPU3.4GHz, GTX1060, windows 10. the The public MSR_Action3D is selected as the experiment dataset. The database has 557 depth image samples and 20 different actions including high wave (A01), horizontal wave (A02), throw (A03), right hand grasp (A04), punch (A05), high throw (A06), cross (A07), hook (A08), circle (A08)(A09), clap (A10), hand swing (A11), side jab (A12), bend (A13), front kick (A14), side kick (A15), jog (A16), tennis swing (A17), tennis serve (A18), golf swing (A19), pick up throw (A20). Ten persons participate in the experiment, each person conducts action 2 to 3 times. In this paper, the original depth map sequence is called a positive order action, marked as Data1. The inverse action is called negative order action. In this paper, the inverse action is obtained by the reverse order of positive order action. The combination of the positive order action and inverse order action forms Data2. The positive order action in dataset 2 is the same as that in Data1. Inverse action contains inverse high wave (B01), inverse horizontal wave (B02), inverse throwing (B03), inverse grasp (B04), inverse strike (B05), inverse high throw (B06), inverse fork (B07), inverse hook (B08), inverse circle (B09), inverse clap (B10), inverse hands up swing (B11), inverse side jab (B12), inverse bend (B13), inverse forward kick (B14), inverse side kick (B15), inverse jog (B16), inverse tennis swing (B17), inverse tennis serve (B18), inverse golf swing (B19), inverse pick up throw (B20).

4.1. Experiment Set

Setting 1. Divide the actions in the data set into 3 groups, the actions with high similarity in the same group. The actions in Data1 are divided into AS1, AS2 and AS3. The actions in Data2 are classified as AS4, AS5 and AS6. The grouping of Data1 and Data2 is shown in table 1. Each group is tested three times. In test 1, 1/3 samples are used as training data, and the remaining samples are as test data. In test 2, 1/2 samples are used as training data, and the remaining samples are as test data. In test 3, 2/3 samples are used as training data, and the remaining samples are as test data.

Setting 2. Cross-verify is conducted for the whole action in the dataset. The samples are divided into 5 parts, where 4 parts are used for training and 1 part is used for testing. The final recognition result is the average of the five results. In this paper, the image block is 10×10 pixel. HOG feature is extracted by sliding image block with step size of 10 pixels. The local binary pattern features of the image are extracted by setting the parameters with a sampling radius of 2 and a sampling number of 8.

Setting 3. Ablation experiments are conducted to show the effectiveness of proposed method in terms of the addition of Depth Sequence Key-frames and Discriminative Collaborative Representation Classifier on the Data1 and Data2.

Table 1. Subsets of Data1 and Data2

Data1	Data1	Data1	Data2	Data2	Data2
AS1	AS2	AS3	AS4	AS5	AS6
A02	A01	A06	A02+B02	A01+B01	A06+B06
A03	A04	A14	A03+B03	A04+B04	A14+B14
A05	A07	A15	A05+B05	A07+B07	A15+B15
A06	A08	A16	A06+B06	A08+B08	A16+B16
A10	A09	A17	A10+B10	A09+B09	A17+B17
A13	A11	A18	A13+B13	A11+B11	A18+B18
A18	A12	A19	A18+B18	A12+B12	A19+B19
A20	A14	A20	A20+B20	A14+B14	A20+B20

Setting 4. The confusion matrix experiment results show that the unique complete timing of DSTEFE plays an important role in human behavior recognition on the database with both positive and reverse sequence behaviors.

4.2. Recognition results

According to the setting 1, the DSTEFE-HOG feature of each action in the three sub-datasets of Data1 is input into different classifiers for classification. In this paper, we select the four classical classifiers including Gaussian Bayes, Random Forest, K-nearest neighbor, SVM to make comparison. Next work direction is research more classifiers. Table 2 displays the recognition results of DSTEFE-HOG features in different classifiers. Table 3 shows the ablation experiment result.

Table 2. Recognition results of DSTEFE-HOG in different classifiers

Classifier	AS1	AS2	AS3
Gaussian Bayes	79.87	76.53	85.39
Random Forest	85.12	84.79	91.96
K-nearest neighbor	81.34	85.72	82.56
SVM	93.77	92.56	96.14
DCRC	98.91	95.71	99.25

Table 3. Recognition results of ablation experiment

Sub-model	AS1	AS2	AS3
Depth Sequence Key-frame	66.78	75.36	81.28
Depth Sequence Key-frame+DCRC	98.91	95.71	99.25

From table 2, it can be seen that DSTEFE-HOG has a high recognition accuracy on all classifiers, but the proposed DCRC has the best classification effect. In order to achieve the most ideal recognition effect for DSTEFE-HOG feature, DCRC is used as the classifier in the following experiments. Table 3 shows that through our proposed scheme, the results of human behavior recognition have been greatly improved.

When key frame algorithm is carried out, the number of key frames N must be determined first. N directly affects the extraction speed of feature map and the removal of redundant information. Figure 7 shows the DSTEFE of the width axis of the tennis swing with different N. Figure 7(a) is the DSTEFE without key frame algorithm. It can be clearly seen from the contents in the white box that the feature map contains more redundant information. In figure 7(b), the number of key frames is 40, and many depth frames still belong to redundant frames, so the effect improvement is not ideal. In figure 7(d), the number of key frames is 25, which clearly shows that the depth frames of many key positions are lost, resulting in inaccurate action description. In figure 7(c), the number of key frames is 30, which not only eliminates redundant information in depth map sequence, but also retains the key information completely.

In this paper, in order to obtain the most ideal key frame sequence, the step length is set as 5 frames, and the recognition accuracy of the final extracted DSTEFE-HOG is taken as the standard value to find the most appropriate key frame number N from 25$\tilde{4}$0 frames. According to setting 1, each action in the three sub-datasets of Data1 is extracted by key frame. The DSTEFE-HOG feature is calculated and results are shown in figure 8. Through the analysis of figure 8, it can be seen more intuitively that when N=30, the recognition accuracy on any sub-datasets is the highest, which indicates that the key frame sequence can best describe the depth map sequence when N=30.

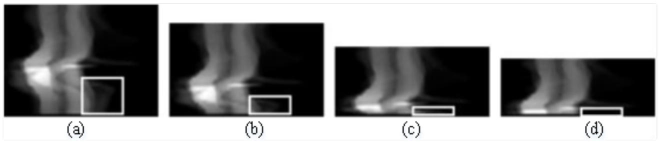


Fig. 7. DSTEFE effect with different key frames. (a) without key frame; (b) N=40; (c) N=30; (d) N=25.

In order to further verify the effectiveness of the key frame algorithm, this paper carries out a comparison experiment on the action recognition effect with/without the key frame algorithm on three sub-datasets according to experiment setting 1. Data1 contains 20 positive order actions. The results of different methods without and with key frame algorithm are shown in table 4 and table 5, respectively.

The recognition comparison results before and after the key frame algorithm in table 4 and table 5 show that the key frame algorithm eliminates the redundant frames in the depth map sequence, reduces the redundant information in the sequence, and improves the final recognition accuracy. Where, the recognition results of DSTEFE-HOG feature processed by key frame are significantly improved compared with those without key frame. The reason is that DSTEFE is formed by projecting the energy information of each frame in depth map sequence onto three orthogonal axes and spliced in time order, which is sensitive to redundant information. The key frame algorithm eliminates redundant frames,

Table 4. Recognition rate of different methods on Data1 without key frame algorithm (%)

Data	AS1	AS1	AS1	AS2	AS2	AS2	AS3	AS3	AS3
Method	Test1	Test2	Test3	Test1	Test2	Test3	Test1	Test2	Test3
MEI+HOG	73.41	86.35	86.41	73.14	81.69	86.95	72.41	71.28	90.65
MEI+LBP	56.27	65.24	71.34	53.40	62.39	73.79	54.84	60.47	75.79
MHI+HOG	69.97	83.60	86.42	64.58	81.69	88.27	72.99	72.18	90.65
MHI+LBP	53.53	66.72	68.61	56.01	65.02	71.16	54.84	54.16	71.73
DMM+HOG	76.14	84.51	87.78	71.82	85.21	86.95	77.81	75.79	82.67
DMM+LBP	57.64	75.34	86.41	63.93	71.16	78.99	64.97	66.87	83.89
DSTEFE+HOG	91.24	89.74	94.25	79.78	86.17	90.24	81.38	84.56	94.77

Table 5. Recognition rate of different methods on Data2 with key frame algorithm (%)

Data	AS1	AS1	AS1	AS2	AS2	AS2	AS3	AS3	AS3
Method	Test1	Test2	Test3	Test1	Test2	Test3	Test1	Test2	Test3
MEI+HOG	74.08	84.51	86.41	74.45	81.69	86.95	75.79	72.18	91.99
MEI+LBP	56.27	65.24	75.45	52.09	58.89	68.53	54.16	59.57	75.79
MHI+HOG	70.67	83.61	86.42	69.19	81.69	88.27	76.46	73.11	90.65
MHI+LBP	53.54	67.08	71.34	56.11	65.12	73.79	55.84	61.37	73.08
DMM+HOG	72.03	88.17	91.89	77.08	86.95	86.27	77.81	77.59	94.71
DMM+LBP	63.81	78.09	86.41	66.56	78.18	85.64	68.35	70.38	87.85
DSTEFE+HOG	91.84	92.19	98.74	85.93	89.98	92.77	89.68	90.34	98.79

reduces redundant information, and makes DSTEFE with the same action have stronger similarity, so the final recognition accuracy is greatly improved. The confusion matrixes of DSTEFE+HOG feature on Data1 and Data2 are shown in figure 8 and 9 respectively.

We also select three other state-of-the-art methods to make comparison containing DHS [15], R-STDP [16] and MMNN [17]. The average results are shown in table 6.

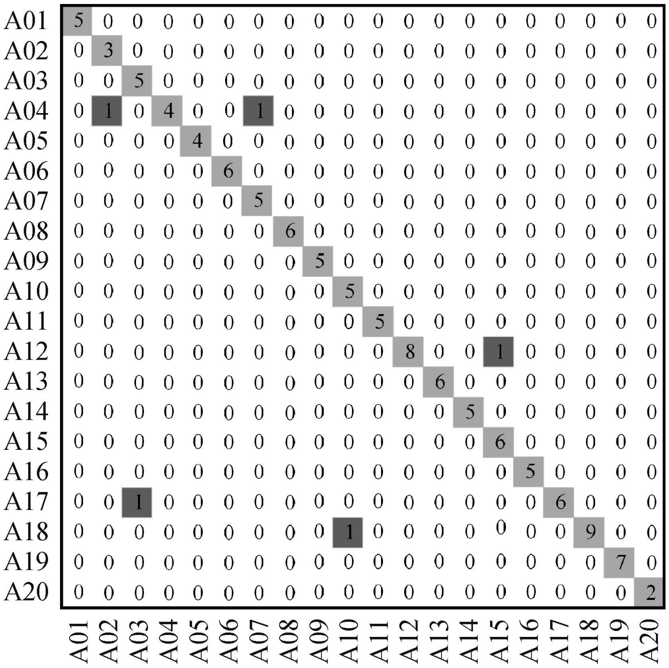
Table 6. Recognition with different methods (%)

Method	Data1	Data2
DHS	76.52	58.17
R-STDP	81.55	79.25
MMNN	93.86	85.44
DSTEFE+DCRC	96.48	91.37

As can be seen from table 6, the recognition accuracy of the proposed method in this paper has been improved compared with other methods.

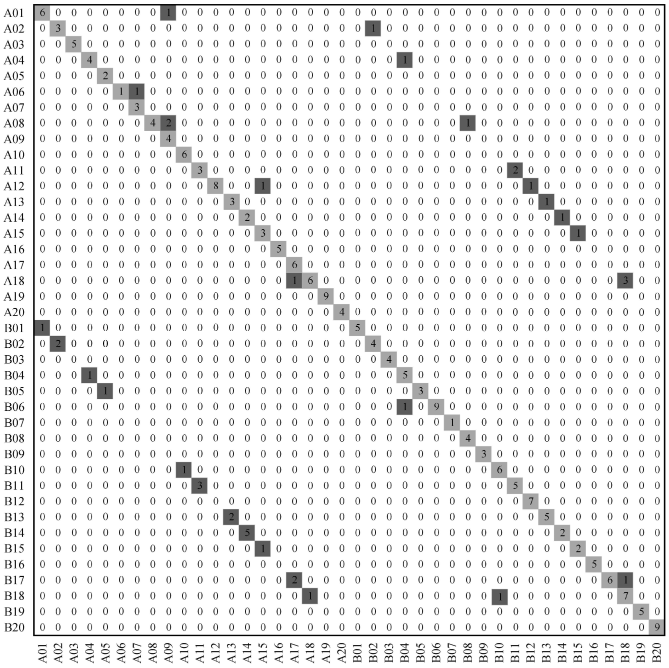
This paper compares the complexity of DSTEFE+DCRC with that of other methods, and the comparison results are shown in Table 7.

In Table 7, f is the frame number of depth map sequence, and the upper limit is 30. W, H and D are width value, height value and depth value of depth map sequence respectively. In this paper, $W = 320$, $H = 240$ and $D = 255$. It can be concluded that the time complexity of DSTEFE+DCRC is lower than that of DHS, R-STDP and MMNN.



	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20
A01	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A02	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A03	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A04	0	1	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
A05	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A06	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A07	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
A08	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
A09	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
A10	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
A11	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
A12	0	0	0	0	0	0	0	0	0	0	0	8	0	0	1	0	0	0	0	0
A13	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0
A14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
A15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
A16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
A17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
A18	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	9	0	0
A19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0
A20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Fig. 8. Confusion matrix of DSTEFE+HOG on Data1.



	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
A01	6	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A02	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A03	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A04	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A05	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A06	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A07	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A08	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
A09	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A10	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A11	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
A12	0	0	0	0	0	0	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
A13	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
A14	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
A15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
A16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
A19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B01	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B02	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B04	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B05	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
B08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
B09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
B10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
B11	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
B12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
B13	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
B14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
B15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
B16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
B17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	0	0
B18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	0	0
B19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
B20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9

Fig. 9. Confusion matrix of DSTEFE+HOG on Data2.

Table 7. Comparison of computational complexity with different methods (%)

Method	Time complexity
DHS	$O(wh) + O(fdhw)$
R-STDP	$O(wh) + O(fwh)$
MMNN	$O(fwh) + O[(f-1)(wh + wd + hd)]$
DSTEFE+DCRC	$O(fwh) + O(fh + fd + wh)$

5. Conclusions

In this paper, we propose a new deep map sequence feature expression method based on discriminative collaborative representation classifier. It solves the problem of missing time sequence information in feature map generated from deep map sequence. The experiment results show that the key frame algorithm improves the extraction rate of feature map and the recognition accuracy of human action. DSTEFE+DCRC not only preserves high recognition accuracy on the positive order action data, but also maintains high recognition accuracy on the inverse order action data. In the future, we will continue to research the action recognition based on deep learning methods and apply them into the practical engineering applications.

Acknowledgments. Harbin Federation of Social Science research project, project title ”In the ”post-Winter Olympics” era, the ice and snow industry was taken as the endogenous driving force to optimize the industrial aggregation mode in Longjiang”.

Availability of data and materials. The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests. The authors declare that they have no conflicts of interest.

References

1. Berlin S J, John M. ”R-STDP Based Spiking Neural Network for Human Action Recognition,” *Applied Artificial Intelligence*, vol. 3, pp. 1-18, 2020.
2. Jisi A and Shoulin Yin. ”A New Feature Fusion Network for Student Behavior Recognition in Education,” *Journal of Applied Science and Engineering*, vol. 24, no. 2, 2021.
3. Bobick A F, Davis J W. ”The recognition of human movement using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.
4. S. Yin and H. Li. Hot ”Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5862-5871, 2020. doi: 10.1109/JSTARS.2020.3025582.
5. Y. Zhu, W. Chen and G. Guo. ”Fusing Spatiotemporal Features and Joints for 3D Action Recognition,” *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland*, OR, pp. 486-491, 2013, doi: 10.1109/CVPRW.2013.78.
6. Luo. ”Feature Extraction and Recognition for Human Action Recognition,” *Machine Vision & Applications*, vol. 25, no. 7, pp. 1793-1812, 2014.

7. Xuan Son Nguyen, Thanh Phuong Nguyen and F. Charpillet. ”Improving surface normals based action recognition in depth images,” *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs*, CO, pp. 109-114, 2016. doi: 10.1109/AVSS.2016.7738053.
8. Q. Nie, J. Wang, X. Wang and Y. Liu. ”View-Invariant Human Action Recognition Based on a 3D Bio-Constrained Skeleton Model,” *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3959-3972, Aug. 2019. doi: 10.1109/TIP.2019.2907048.
9. S. Chaudhary and S. Murala. ”Depth-based end-to-end deep network for human action recognition,” *IET Computer Vision*, vol. 13, no. 1, pp. 15-22, 2019. doi: 10.1049/iet-cvi.2018.5020.
10. Mattiev, J., Kavek, B. ”Distance based Clustering of Class Association Rules to Build a Compact, Accurate and Descriptive Classifier,” *Computer Science and Information Systems*, Vol. 18, No. 3, pp. 791-811. (2021), https://doi.org/10.2298/CSIS200430037M.
11. Fan, Z., Guan, Y. ”Face Recognition Based on Full Convolutional Neural Network Based on Transfer Learning Model,” *Computer Science and Information Systems*, Vol. 18, No. 4, pp. 1395-1409. (2021), https://doi.org/10.2298/CSIS200922028F.
12. Chao X, Hou ZJ, Li X, Liang JZ, Huan J and Liu H Y. ”Action recognition under depth spatial-temporal energy feature representation,” *Journal of Image and Graphics*, vol. 25, no. 04, pp. 0836-0850, 2020.
13. Yang X, Zhang C, Tian Y L. ”Recognizing actions using depth motion maps-based histograms of oriented gradients,” *ACM International Conference on Multimedia. ACM*, 2012:1057.
14. S. Jia, L. Shen and Q. Li. ”Gabor Feature-Based Collaborative Representation for Hyperspectral Imagery Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 1118-1129, Feb. 2015. doi: 10.1109/TGRS.2014.2334608.
15. Baofeng Z, Jun K, Min J. ”Human Action Recognition Based on Discriminative Collaborative Representation Classifier,” *Laser & Optoelectronics Progress*, vol. 55, no. 1, pp. 257-263, 2018.
16. Md Azher, Uddin, Young-Koo, et al. ”Feature Fusion of Deep Spatial Features and Handcrafted Spatiotemporal Features for Human Action Recognition,” *Sensors*, vol. 19, no. 7, pp. 1599, 2019. doi: 10.3390/s19071599.
17. Berlin S J, John M. ”R-STDP Based Spiking Neural Network for Human Action Recognition,” *Applied Artificial Intelligence*, vol. 3, pp. 1-18, 2020. Zhao H, Xue W, Li X, et al. ”Multi-Mode Neural Network for Human Action Recognition,” *IET Computer Vision*, vol. 14, no. 8, pp. 587-596, 2020.

Wang Yuhang is an associate professor in School of Physical Education, Harbin University. His current research interests include physical education, and behavior analysis.

Tao Feng is a lecturer in Physical Education Department of Harbin Institute of Finance, PhD candidate at The Graduate School Saint Paul University. His research interests include video quality of service over wireless networks, adaptation, perceptual modeling, physical education, and behavior analysis.

Yi Zheng is the director of educational affairs office of Physical Education Institute of Harbin University. His current research interests include Machining and Modeling of physical education, and behavior analysis.

Received: March 21, 2022; Accepted: August 28, 2022.