Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 8780105, 10 pages
https://doi.org/10.1155/2018/8780105

Research Article

Action Recognition Based on Depth Motion Map and Hybrid Classifier

Wenhui Li, Qiuling Wang, and Ying Wang

College of Computer Science and Technology, Jilin University, Changchun 130012, China

Correspondence should be addressed to Ying Wang; wangying_jlu@163.com

Received 4 July 2018; Revised 31 October 2018; Accepted 1 November 2018; Published 14 November 2018

Academic Editor: Nazrul Islam

In order to efficiently extract and encode 3D information of human action from depth images, we present a feature extraction and recognition method based on depth video sequences. First, depth images are projected continuously onto three planes of Cartesian coordinate system, and differential images of the respective projection surfaces are accumulated to obtain the complete 3D information of the depth motion maps (DMMs). Then, discriminative completed LBP (disCLBP) encodes depth motion maps to extract effective human action information. A hybrid classifier combined with Extreme Learning Machine (ELM) and collaborative representation classification (CRC) is employed to reduce the computational complexity while reducing the impact of noise. The proposed method is tested on the MSR-Action3D database; the experimental results show that it achieves 96.0% accuracy and well performs better robustness comparing to other popular approaches.

1. Introduction

Human action recognition is an important and challenging topic in the field of computer vision. Early action recognition uses traditional color camera to capture video sequences [1]. Action recognition methods based on traditional color camera data are usually divided into two major categories. The first class is to directly classify action features without considering temporal information. For example, in [2], authors propose a video-based contour extraction method to extract the human action contour map from videos, then Hu Moment as the distance measurement is applied to represent the distance between the motion observation sequences and the training data. In [3], action recognition using temporal gradients, optical flows, and Support Vector Machine (SVM) is proposed. The second class takes into consideration temporal and spatial features for classification. In [4], authors use the key gestures of human action to establish an HMM model and train dynamic information according to the model. In [5], a hybrid hidden Markov model is utilized to solve multiview problems, and it combines shape and motion optical flow to classify motions. This method has the advantages of simple calculation and getting parameters such as human position and size of human appearance easily.

However, problem which cannot be solved by the traditional methods is that color camera cannot capture spatial information whereas human actions occur in 3D space. Moreover, the video data could be affected by many factors such as illumination conditions and background variation, then the processing algorithm cannot obtain high recognition performance due to lack of spatial information.

In recent years, depth cameras have become increasingly popular. Devices such as Kinect or ASUS Xtion can acquire RGB, depth, and skeletal data streams simultaneously in real time. The rich spatial information has opened new research directions for human action recognition research. Researchers have carried out many studies on human behavior recognition based on 3D skeletal nodes and depth images. For example, Fitzgibbon [6] uses the Microsoft somatosensory device Kinect to propose a method for estimating the position of skeletal joints by extracting the shape information of human motion. Subsequently, in [7], the authors apply the algorithm to the opponent's pose estimation. They use random forest to classify the pixel points and then estimate the position of the joint in the hand. In [8], authors establish a spherical coordinate system which is independent of the angle of view based on the skeletal data, and they ignore the difference in body size between humans; LDA is applied

to reduce the dimension of feature vectors. K-Means is used for clustering. Finally, they use discrete hidden Markov model to do the classification. In [9], motion information and pose estimation extracted using Kinect are combined to construct the Eigen Joints description, and then Naive Bayes Classifier is employed to recognize human action. In [10], authors use Local Binary Pattern and Extreme Learning Machine to identify human action. This approach achieves well performance on the testing data set.

From the existed approaches, we can find that elaborate feature and well-designed classifier play the critical role in performance improvement. Using skeletal data as discriminated feature will be limited by the inaccuracy of skeletal joint's position. Although the calculation is simple, the recognition rate is relatively low. Using the original depth image data for action recognition has a higher recognition rate, but the redundancy of the data features will increase the time complexity.

In order to obtain the trade-off between recognition accuracy and computational complexity, in this paper, we propose a framework for human action recognition. Our method is using discriminative completed local binary pattern based on depth motion maps as feature and is using a hybrid classifier combined with Extreme Learning Machine (ELM) and collaborative representation classification (CRC) to finish the classification. The proposed method has been tested on MSR-Action3D database, in which there is a single person in each frame of sequence. The experiment results show that the proposed algorithm has a high recognition rate and good robustness.

The rest of this article is structured as follows: Section 2 describes the features of depth motion maps and the discriminative completed local binary pattern algorithm based on depth motion maps features. Section 3 introduces the proposed ELM-CRC hybrid classifier and, at the same time, gives the specific implementation principle. Section 4 supplies the experiment setting and results analysis. Section 5 summarizes the paper.

2. Description of Motion Features

2.1. Depth Motion Feature. The concept of DMMs was proposed in [18]. In order to make full use of the 3D structure and shape information of depth images, depth data of each human action frame is projected onto three orthogonal Cartesian planes, respectively, namely, front view, the top view, and the left view. In order to reduce the computational complexity, each projection view consisting of successive depth data frame differences is modified on the above basis [12]. Each frame action can be expressed as v = {f, s, t}, where f, s, t are the projections of the depth data difference in the front view, left view, and top view, respectively. The depth motion maps DMM_v feature is calculated from

$$DMM_v = \sum_{i=1}^{N-1} \left| map_v^{i+1} - map_v^i \right| \tag{1}$$

where i is the ith frame in time series and N is the total number of video sequence frames.

For much more computational efficiency, we use the methods in [12] to generate DMM_f, DMM_s, and DMM_t. That is to keep the image size consistent and extract the region of interest. The image is cropped to remove background points in the depth sequence frame. Then, the final DMM is the results after foreground extraction.

2.2. DMMs-Based disCLBP Features. Traditional local binary pattern (LBP) [19] is an effective feature extraction algorithm that has been widely used in various applications [20]. For a gray image, the original LBP operator is defined as

$$LBP_{P,R} = \sum_{p=0}^{p-1} s\left(g_p - g_c\right) 2^p,$$
$$S(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x \leq 0 \end{cases} \tag{2}$$

where g_c is the gray value at the center of the window(x_c, y_c) and g_p is the gray value of the P field dot uniformly distributed on the circumference of the center point (x_c, y_c) with radius R.

In order to consider the difference of brightness and amplitude between the center pixel and the neighborhood pixels, Guo[21] proposed a completed local binary pattern CLBP operator. A local region is represented by the central pixel and the local difference (LD) sign-magnitude transform (LDSMT). He proposed three different descriptors: the center descriptor (CLBP-Center, CLBP_C), the sign descriptor (CLBP-Sign, CLBP_S), and the magnitude descriptor (CLBP-Magnitude, CLBP_M). The feature extraction process is shown in Figure 1.

The difference between the gray value of the center pixel and the adjacent region is expressed as $d_p = s_p \times m_p$, s_p is the sign of d_p, and m_p is the absolute value of d_p. Here, s_p is the encoding rule for the descriptor CLBP_S: if $s_p \geq 0$, then CLBP_S is 1; otherwise, it is -1. The remaining two descriptors CLBP_M, CLBP_C are calculated as

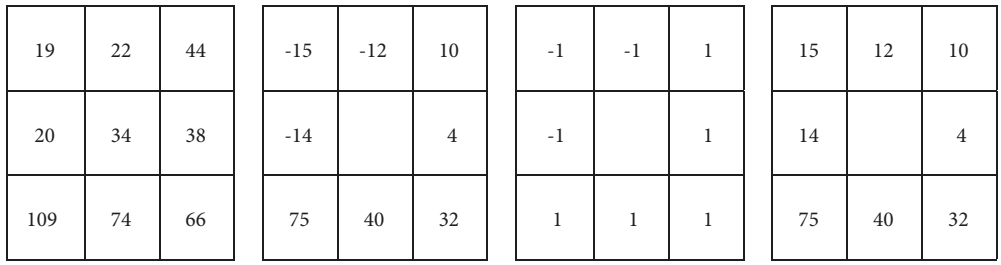
$$CLBP\_M_{P,R} = \sum_{p=0}^{p-1} s\left(m_p, c\right) 2^p \tag{3}$$

$$CLBP\_C_{P,R} = s\left(g_p, c\right) \tag{4}$$

$$\mathrm{s}(a, c) = \begin{cases} 1, & a \geq c \\ 0, & a < c \end{cases} \tag{5}$$

where $c$is the mean of m_p in the local patch.

The feature information extracted by CLBP operator is more comprehensive, but the feature dimension is also increased, which brings more time consumption. In order to reduce the dimension the operators (LBP and its extended versions) and to select more robust features, [22] adopts local-global training strategy based on all LBP models which is called disCLBP. By using the method considering the smallest intraclass distance and the largest interclass distance, features with strong classification ability are selected. To extract the discriminated features of the depth images and to ensure time



(a) 3 × 3 image (b) LD image (c) CLBP_S (d) CLBP_M

Figure 1: CLBP operator.

efficiency, we use disCLBP algorithm as the feature extraction method.

Suppose the training depth images contain J classes, and each class has S depth images. Count LBP pattern sets with probability greater than a certain threshold in each image of each class, then these patterns will be as the characteristic representation of the image. In this way, features with less contribution in the sample features set are removed, so that a LBP mode sets with S patterns are obtained, as shown in

$$J_i = \arg\min_{|J_i|} \left(\frac{\sum_{j \in J_i} f_{i,j}}{\sum_{k=1}^{p} f_{i,k}} \right) \geq n\%, \quad J_i \subseteq [1, 2, \ldots, p] \tag{6}$$

where J_i is the set of selected feature types, $|J_i|$ is the number of elements in the set J_i, p is the total number of original mode types, and $f_{i,j}$ is the feature value of the jth mode type of the picture i.

For all the depth images of the same class, the intersection of the LBP feature patterns of all the images is used as the dominant LBP features of this class. As shown in Figure 2, the selected common features of the three depth images are P5, P8.

Then, the union of the dominant LBP features of all classes is the disCLBP, which is the global LBP feature dictionary of all depth images. The LBP feature description of each image is the histogram distribution of disCLBP in each image.

Taking clapping hand as an example, Figure 3(a) is the depth image sequence of clapping, Figure 3(b) is the image obtained by projecting a depth image from the front view over a period of time, and Figure 3(c) is the preprocessing results which normalizes the images in Figure 3(b) to avoid the problem of computational complexity caused by the background regions.

3. Hybrid Classifier Based on ELM and CRC

3.1. ELM. Comparing with the traditional neural network, ELM has the advantages of fast training and well generalization performance [18]. ELM is originally proposed for the Single-hidden Layer Feedforward Neural-networks (SLFNs). Then, it extends to the generalized feedforward network. The significant advantage of ELM is that the model is a random parameter model. In ELM network, the input layer weights and the biases are random, and only the output weights need to be determined.

Suppose there are N samples (x_j, t_j), j $= 1, 2, \ldots, N$, where input vector $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})^T \in R^n$, expected output $t_j = (t_{j1}, t_{j2}, \ldots, t_{jm})^T \in R^m$. For a SLFNs with L hidden nodes, the ELM can be represented as

$$\sum_{i=1}^{L} \beta_i g\left(a_i \bullet x_j + b_i\right) = o_j \tag{7}$$

where $g(x)$ is the activation function; the learning goal of the SLFNs is to minimize the output error which is expressed as

$$\sum_{j=1}^{N} \left\| o_j - t_j \right\| = 0 \tag{8}$$

That is, there are β_i, a_i, and b_i, making (9) true:

$$\sum_{i=1}^{L} \beta_i g\left(a_i \bullet x_j + b_i\right) = t_j, \quad j = 1, 2, \ldots, N \tag{9}$$

If H represents the output of the hidden node, β is the output weight matrix, T is the expected output, and (9) can be expressed as

$$\mathrm{H}\beta = \mathrm{T} \tag{10}$$

where $\mathrm{H} = \begin{bmatrix} g(a_1 \bullet x_1 + b_1) & \cdots & g(a_L \bullet x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(a_1 \bullet x_N + b_1) & \cdots & g(a_L \bullet x_N + b_L) \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}$,

$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}$

In the SLFNs learning machine, once the input weight and hidden layer bias are randomly set, it will uniquely determine the output weight matrix of the hidden layer. Therefore, training SLFNs learning machine can be converted into solving the linear matrix equation $\mathrm{H}\beta = \mathrm{T}$, and the output weight matrix β can be determined by

$$\hat{\beta} = H^{+}\mathrm{T} \tag{11}$$

In (11), H^+ is the Moore-Penrose generalized inverse of H. Since ELM solves the output using the classical Least Squares

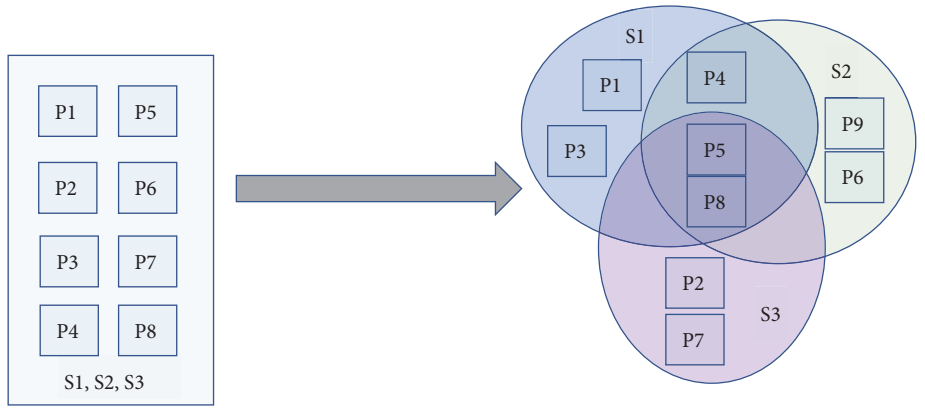


Figure 2: Extract public features.

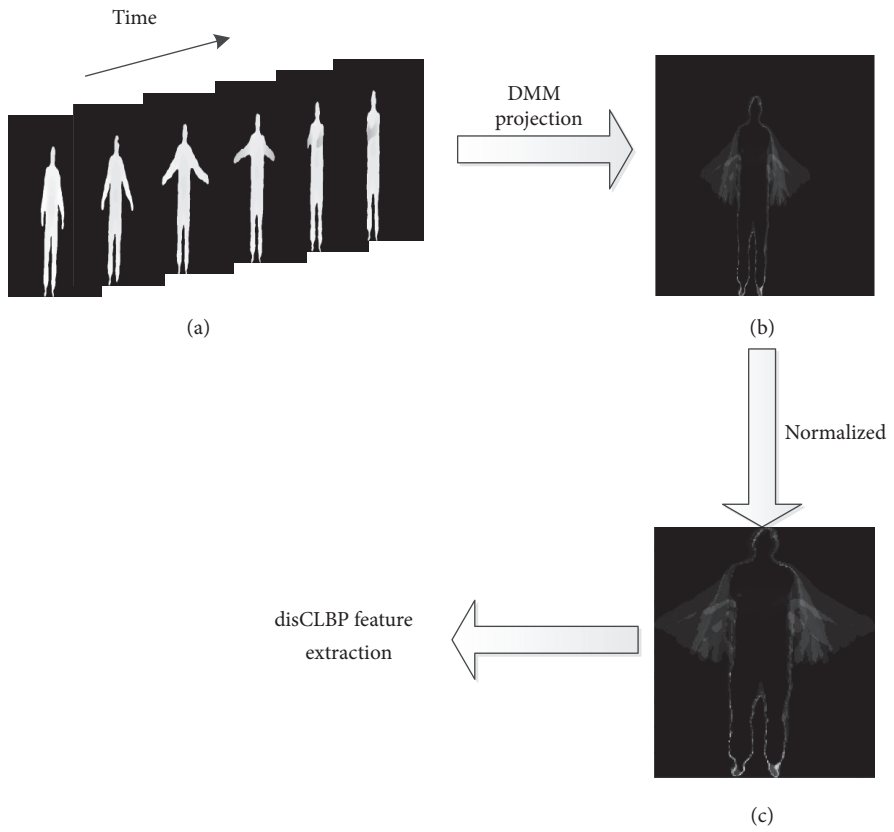


Figure 3: Depth motion maps feature formation process.

Method, it is prone to get singular values and the results are unstable. Therefore, the improved regularized Extreme Learning Machine is used in this paper as

$$\widehat{\beta} = H^{+}\mathrm{T} = H^{T}\left(\frac{I}{C} + HH^{T}\right)^{-1} T \tag{12}$$

where I is a diagonal matrix and C is a regularization coefficient.

3.2. CRC. In 2009, John Wright and other scholars proposed a face recognition method based on sparse representation-based classification (SRC) [23]. When face images are subject to noise pollution or other error interference, they still obtain better recognition.

The training data set can be represented as $D = [D_1, D_2, \ldots, D_c] \in R^{m\times n}$, where $D_i = [d_{i1}, d_{i2}, \ldots, d_{in}] \in R^{m\times ni}$ is a matrix composed of the ith class training image vectors, d_{ij} is the jth training image vector of the class i, c is the number of image class, m is the dimension of the training

sample image vector, and n is the number of training sample images. The matrix D is used as a dictionary, for the test sample y, it can be expressed as $y = Dx$, where x is the sparse representation vector of sample y under the dictionary D. Sparse representation classification solves the problem of minimizing l_1 norm:

$$\begin{aligned} \widehat{\alpha} &= \arg\min_{\alpha} \quad \|\alpha\|_1 \\ s.t. &\quad \|y - D\alpha\| \le \varepsilon \end{aligned} \tag{13}$$

where ε is the noise in y. The matrix dimension used to describe the image error and noise in the SRC is too high, so that the computational complexity is very large.

The collaborative representation-based classification [24] (CRC) method uses a weaker sparse l_2 norm to convert the minimization l_1 norm constraint problem into the least squares constraint problem. It uses a regularized mean squared error method, as shown in (14), where I is the unit matrix. The sparseness of α is much smaller than the sparseness of the l_1 norm constraint. It can be easily and quickly calculated by (15), which greatly improves the computational efficiency and computation speed. In this paper, we use CRC method for classification.

$$\widehat{\alpha} = \arg\min_{\alpha} \left(\|y - D\alpha\|_2^2 + \lambda \|\alpha\|_1 \right) \tag{14}$$

$$\widehat{\alpha} = \left(D^T D + \lambda I \right)^{-1} D^T y \tag{15}$$

3.3. ELM and CRC Hybrid Classifier. ELM only needs to determine the network structure parameters; it can simulate the potential law between input and output. It does not need to adjust parameters when classifying features, so that it can perform fast parallel operations. CRC can obtain sparse coefficients through many iterations. Although there are many methods which can quickly solve the sparse coefficients, but compared with ELM, the processing speed is still very slow. However, when the images are classified, CRC still has strong robustness under the conditions of occlusion and illumination, but ELM does not have this advantage.

To make full use of the ELM fast training or testing and the superior ability of CRC to select feature insensitive when processing noise images, in this paper, we use the ELM-CRC hybrid classifier. And we propose an estimation criterion for misclassified images in ELM and an algorithm for adaptively reducing dictionary dimensions. It can adaptively select ELM and CRC to accelerate the entire classification process. The low-noise image is processed by ELM; the noise image is processed by CRC. Thus, the ELM misclassified image can be processed by a robust CRC classifier.

For a sample data belonging to class p, under the +1, -1 output coding, the standard ELM expected output is $t = (-1, -1, \ldots, -1, 1, -1, \ldots, -1)^T$, and 1 is the pth element. Assume that the real output is $o = (-1, o_f, -1, \ldots, -1, o_s, -1, \ldots, -1)^T$, where o_f, o_s are the first and second largest values of the vector o. If the training errors of the ELM classifier are minimal; in the ideal case, the desired output t and the real output o should satisfy the relationship: $t \approx o$, thus we can get $o_f - o_s \approx t_f - t_s = 2$. But in general, image data generally has some noise. ELM with zero train error will reduce the generalization ability of the network.

In this paper, we select 275 depth images from the database for ELM test, including 20 actions performed by individuals 1, 3, 5, 7, and 9. Figure 4 shows the distribution of misclassified samples with respect to $o_f - o_s$. It can be seen that (1) when $o_f - o_s < 0.1$, half of the corresponding samples are misclassified by the ELM classifier; (2) all misclassified images almost all satisfy $o_f - o_s < 0.7$. The approximate judgment of a misclassified image by the ELM can be described as filtering of the noise images.

We use $\text{Tdiff} = o_f - o_s$ to compare with a threshold σ for noise image discrimination, where o_f and o_s represent the first and second largest entry in the ELM output vector. If $\text{Tdiff} > \sigma$, the classification of test image is determined by ELM. Otherwise, the test image will be provided to the CRC. In general, the larger the Tdiff, the better the classification effect[25].

In [24], authors point out that using general and overcomplete dictionaries to query image is lack of adaptability, due to the negative influence of unrelated classes. So, in the CRC classification stage, we classify the image by including subcategories of similar classes rather than the entire dictionary. We consider the top k elements of the ELM output, because unrelated classes tend to have small responses in the ELM output. Specifically, for the query image y, we record the indexes of the k largest elements in the output vector. Then, we select a train data set having the same label as the k indexes and adaptively construct a subdictionary for collaborative representation. Taking forward punch (action 5) as an example, the desired output and the actual output of the ELM classifier are described in Figure 5. The ELM misclassifies this action as hammer (action 3), so we use the image features corresponding to the k maximum values of the actual output vector for dictionary dimensionality reduction.

Compact subdictionaries are denoted as $A_y^* = A_{m(1)}, A_{m(2)}, \ldots, A_{m(k)}$, where $\text{m(i)} \in \{1, 2, \ldots, \text{k}\}$ is one of the indexes of the k largest entries, and $A_{m(i)}$ represents all the training samples belonging to the m(i) class. Therefore, instead of calculating the sparse representation coefficients over all the training samples, we solve the following problems:

$$\widehat{x} = \left(A_y^{*T} A_y^* + \lambda I \right)^{-1} A_y^{*T} y \tag{16}$$

where I is unit matrix. Finally, we get the class of test sample by

$$\text{Label}(\text{y}) = \arg\min_{\widehat{x}} \left(\left\| y - A_y^* \widehat{x} \right\|_2^2 \right) \tag{17}$$

The hybrid classifier algorithm proposed in this paper is as shown in Algorithm 1.

The algorithm flowchart in this paper is shown in Figure 6; the computational complexity of the proposed method is $O(n + n^2)$.

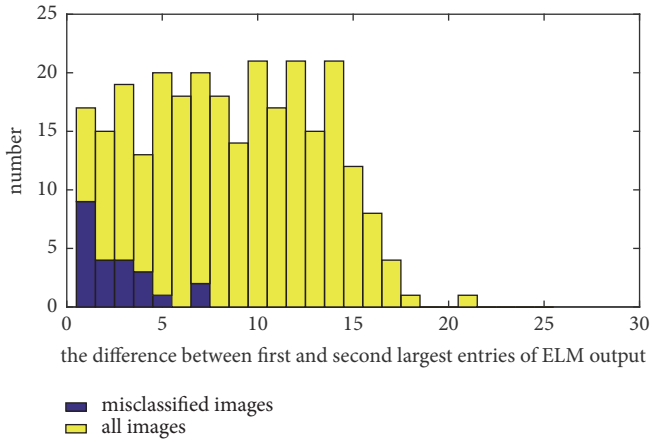


Figure 4: Misclassification samples on $o_f - o_s$ distribution.

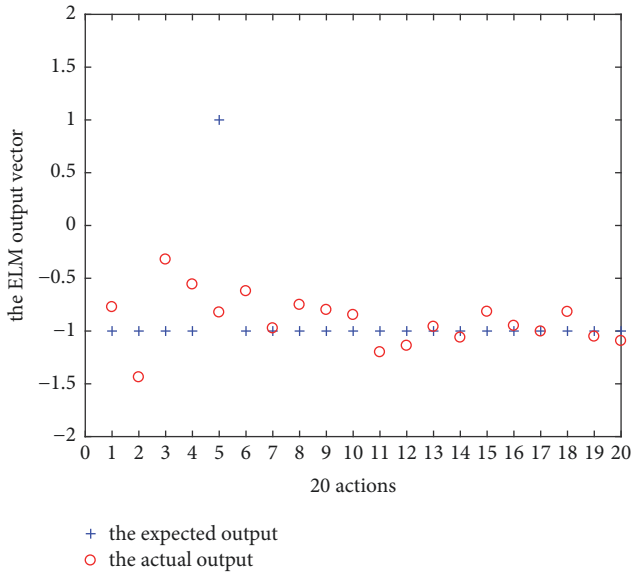


Figure 5: Expected output and actual output of the forward punch.

4. Experimental Results and Analysis

4.1. Experiment Data. The proposed algorithm is tested on MSR-Action3D [13] database, which is an action recognition library for Kinect depth camera. It contains 557 silhouette images of 240×320 resolution and has 557 skeletal data. There is a ground truth for each image. In our experiments, we use depth image data, including 10 people doing 20 actions with 2~3 times: 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 picking throw. Figure 7 is silhouette images of several actions from MSR-Action3D database. The depth camera data used to support the findings of this study are available from the corresponding author upon request.

4.2. Experiment Setting. Setting one is the same as the setting in [13], in which all actions are divided into 3 groups (AS1, AS2, and AS3). Each group has 8 actions, as shown in Table 1. AS1 and AS2 have similar actions. However, in AS3, there are relatively small and complex actions. Each group performs three experiments. In experiment one, 1/3 of the video data is used as train data, and 2/3 is used as test data. In experiment two, 2/3 of the video data was used as train data and 1/3 was used as test data. In experiment three, data of characters (1, 3, 5, 7, and 9) was used as train data, and data of characters (4, 6, 8, and 10) was used as test data.

Setting two is also designed according to the setting in [13]. It divides all the video data into two parts equally. The train and test videos are collected by different people. In the train and test process, they use half of the video data, respectively.

4.3. Experiment Results and Analysis. In order to classify the image features more precisely, under the setting two, we test training accuracy of the samples under different k and σ. As shown in Figure 8, when k=7 and σ=0.3, the training accuracy of image data features can be achieved 96.0%. And when σ remains 0.3, the value of k increases, but the accuracy remains unchanged at 96.0%. Increasing the value of k, time complexity increases with it, so we chooses k=7, σ=0.3 in this paper. The classification accuracy is represented by $\text{Acc} = 1 - (N(err(o_f)) + N(err(k)))/N$, where $N(err(o_f))$ is the number of samples in which the output of the ELM classifier satisfies $o_f - o_s > \sigma$, but the corresponding class of o_f is not the expected classes. And $N(err(k))$ is the number of samples in which the expected class is not in the class set determined by k, but the output of the ELM classifier satisfies $o_f - o_s \leq \sigma$.

In the setting one, the results of the proposed method and the other state-of-the-art methods are shown in Table 2. Generally speaking, the method in this paper has a high accuracy rate in the three groups of experiments. In the experiment one (1/3 for train, 2/3 for test), the recognition rate is equal to the existing methods. In the experiment two (2/3 for train and 1/3 for test), the recognition rate is slightly higher than the existing methods. In experiment three (1/2 for train, including the action data of 1, 3, 5, 6, 7, and 9 people, and the remaining for test), the average recognition rate is

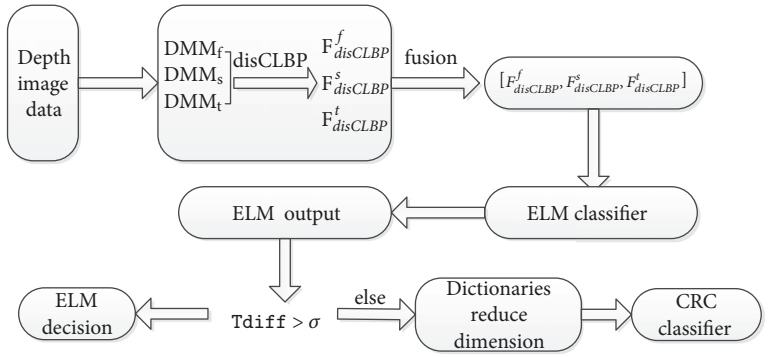


Figure 6: Algorithm flowchart.

Table 1: Setting one.

Action set1(AS1)	Action set2(AS2)	Action set3(AS3)
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

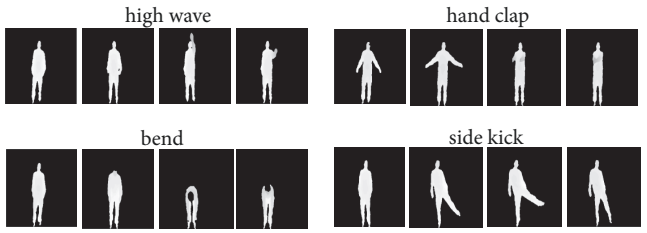


Figure 7: MSR-Action3D database action example.

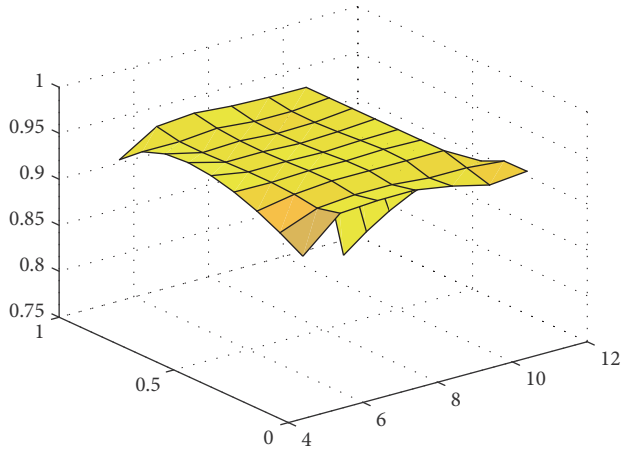


Figure 8: ELM output accuracy varies with k, σ.

significantly higher than the existing methods since there are still large difference between doing the same action by different people. The result of experiment three shows that our method has better robustness.

According to setting two, the test results of this paper and the current existing methods on the MSR-Action3D database are shown in Table 3. Relative setting one, setting two contains more action classes, so it is more challenging than setting one. According to the result, it can be seen that the method in this paper has a high recognition rate of 96.0%.

In order to test the accuracy and time complexity of the ELM and ELM-CRC classifiers, 275 images are tested under the setting two. The experimental environment is matlabR2016a with win10 corei5. Table 4 shows the processing time. It can be seen that ELM classifier has a fast test speed, but its classification accuracy is poor. The ELM-CRC classifier not only has a faster test speed but also has a higher test accuracy.

5. Conclusion

In this paper, we propose a new feature extraction and recognition method based on depth video sequences. By extracting the disCLBP features of the DMMs and using ELM-CRC hybrid classifier, we reduce the computational complexity while reducing the impact of noise data on the classification results. The proposed method is tested on the MSR-Action3D database. Experiment results show that this hybrid classifier not only has better classification accuracy than ELM and the classification speed is very fast. Compared with the current detection algorithms of the same kind, the advantage of this paper is that the errors of the recognition results are better reduced. With the improvement of computer performance, the method proposed in this paper is not constrained by the performance of the computer and can achieve real-time effects. In this paper, the discrimination of similar behaviors

Input: DMMs-based disCLBP features of the test image y; DMMs-based disCLBP features of the training action database A with 20 classes;
Output: The class label of y.
(1) Training the ELM classifier using the action database A;
(2) Calculate the network output o by ELM classifier;
(3) Find the first and second largest entries of ELM output o_f and o_s;
(4) if $o_f - o_s > \sigma$ then
(5) Label(y) = $\arg\max_i o(i)$
(6) else
(7) Set the indexes of k largest entries in o;
(8) Get the sub-dictionary $A_y^* = [A_{m(1)}, A_{m(2)}, \ldots, A_{m(k)}]$ in the action database A;
(9) Solve $\hat{x} = (A_y^{*T} A_y^* + \lambda I)^{-1} A_y^{*T} y$
(10) end if
(11) for i = 1 to k do
(12) Calculate the residuals $r_i = \|y - A_i x_i\|_2^2$;
(13) end for
(14) Label(y) = $\arg\min(r_i)$

Algorithm 1: The proposed ELM-CRC classifier for action recognition.

Table 2: Action recognition results by setting one.

method (%)		Ref.[11]	Ref.[9]	Ref.[12]	Ref.[13]	Our method
experiment one	AS1	97.3	94.7	97.3	89.5	95.3
	AS2	92.2	95.4	96.1	89.0	96.1
	AS3	98.0	97.3	98.7	96.3	98.0
	average	95.8	95.8	97.4	91.6	96.4
experiment two	AS1	98.7	97.3	98.6	93.4	98.6
	AS2	94.7	98.7	98.7	92.9	96.0
	AS3	98.7	97.3	100	96.3	100.0
	average	97.4	97.8	99.1	94.2	98.2
experiment three	AS1	96.2	74.5	96.2	72.9	99.0
	AS2	84.1	96.1	83.2	71.9	90.2
	AS3	94.6	96.4	92.0	79.2	94.6
	average	91.6	89.0	90.5	74.7	94.6

Table 3: Action recognition results by setting two.

method	recognition rate (%)
DMM-HOG[11]	85.5
Eigen Joints[9]	81.4
HON4D[14]	88.36
Actionlet[15]	88.2
Random Occupancy[16]	86.5
Depth Cuboid[17]	91.9
this paper	96.0

Table 4: Experiment results of ELM and ELM-CRC.

method	recognition rate	test time(s)
ELM	92.0%	0.004
ELM-SRC	96.0%	16.52

is lower, and in later work, we consider the extraction of better features to identify.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors wish to acknowledge the support of National Science Foundation of China under Grant U1564211 and Jilin Planned Projects for Science Technology Development under Grant 20170204020GX.

References

[1] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pp. 2004–2011, Miami, FL, June 2009.

[2] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[3] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1253–1260, Rio de Janeiro, Brazil, October 2007.

[4] X. L. Feng and P. Perona, "Human action recognition by sequence of movelet codewords," in *Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission*, pp. 717–721, Padova, Italy, 2002.

[5] M. Ahmad and S.-W. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image sequences," *Pattern Recognition*, vol. 41, no. 7, pp. 2237–2252, 2008.

[6] J. Shotton, T. Sharp, A. Kipman et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[7] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 415–422, Spain, November 2011.

[8] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '12)*, pp. 20–27, Providence, RI, USA, June 2012.

[9] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '12)*, pp. 14–19, Providence, RI, USA, June 2012.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proceedings of the 2015 15th IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, pp. 1092–1099, USA, January 2015.

[11] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia, MM 2012*, pp. 1057–1060, Japan, November 2012.

[12] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-Time Image Processing*, vol. 12, no. 1, pp. 155–163, 2016.

[13] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '10)*, pp. 9–14, San Francisco, Calif, USA, June 2010.

[14] O. Oreifej and Z. Liu, "HON4D: histogram of oriented 4D normals for activity recognition from depth sequences," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 716–723, IEEE, June 2013.

[15] J. Wang, Z. Liu, and Y. Wu, "Learning Actionlet Ensemble for 3D Human Action Recognition," in *Human Action Recognition with Depth Cameras*, SpringerBriefs in Computer Science, pp. 11–40, Springer International Publishing, Cham, 2014.

[16] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part II*, pp. 872–885, Springer, Berlin, Germany, 2012.

[17] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 2834–2841, IEEE, Portland, Ore, USA, June 2013.

[18] Q. Zhu, A. K. Qin, P. N. Suganthan, and G. Huang, "Evolutionary extreme learning machine," *Pattern Recognition*, vol. 38, no. 10, pp. 1759–1763, 2005.

[19] M. Topi, O. Timo, P. Matti, and S. Maricor, "Robust texture classification by subsets of local binary patterns," in *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 3, pp. 935–938, Barcelona, Spain, 2000.

[20] W. Li, C. Chen, H. Su, and Q. Du, "Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3681–3693, 2015.

[21] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[22] Y. Guo, G. Zhao, and M. Pietikäinen, "Discriminative features for texture description," *Pattern Recognition*, vol. 45, no. 10, pp. 3834–3843, 2012.

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, 2009.

[24] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 471–478, Barcelona, Spain, November 2011.

[25] M. Luo and K. Zhang, "A hybrid approach combining extreme learning machine and sparse representation for image classification," *Engineering Applications of Artificial Intelligence*, vol. 27, pp. 228–235, 2014.