Received March 30, 2020, accepted April 4, 2020, date of publication April 7, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986305

Human Motion Target Recognition Using Convolutional Neural Network and Global Constraint Block Matching

WEI LIU[1] AND SHANGBIN LI[2]
[1]Physical Education Department, Heilongjiang University, Harbin 150900, China
[2]Physical Education Department, Harbin Engineering University, Harbin 150001, China
Corresponding author: Shangbin Li (sports@hrbeu.edu.cn)

ABSTRACT The traditional human behavior recognition algorithm is easy to ignore the spatial constraint problem of feature blocks, which leads to poor recognition effect and low correct rate. Therefore, we proposed a human motion target recognition algorithm based on Convolution Neural Network (referred to as the "CNN") and global constraint block matching. First, key frames of the human motion video were extracted, second, the local feature and global feature of key frames were analyzed, and CNN was used to perform feature fusion, then, according to the result of the feature fusion, a feature block was formed and the closest matching feature block is obtained, using the definition of spatial constraint, we considered the spatial data of human motion in the vertical direction, calculates the spatial constraint weight, and further completes the matching. Finally, the score of matching block and the spatial constraint weight were calculated, and the human motion targets are recognized based on the cumulative score. The experimental results show that the proposed algorithm has a high key frame extraction accuracy of more than 90% and less time consumed in feature fusion, high matching accuracy of feature blocks of more than 80%, and high feature blocks, the F-measure of human behavior recognition is 0.95 on average, and the overall recognition performance is good.

INDEX TERMS Convolutional neural network, global constraint block, feature block, spatial constraints, matching, human motion target, recognition.

I. INTRODUCTION

In fact, the problem of human motion target recognition is the acquisition of human features and the reproduction of silhouettes in the multi-camera system. Human motion target recognition belongs to the research field of computer vision [1], [2]. Computer vision is a process used to simulate human vision [3], [4]. It can help humans better understand video information, and it is currently widely used in human-computer interaction, artificial intelligence [5], [6], intelligent video surveillance, and sports, especially sports video analysis [7], [8]. Computer videos usually contain a large amount of human information, and human motion target recognition is an important way to promote computer to understand and master human motion [9], [10], which is the research focus of computer vision. However, the video contains a huge amount of image and data information and has diverse types of environment; due to the characteristics, viewing angle and light, it is relatively difficult for different cameras to accurately recognize the human motion target in actual work [11]. Therefore, how to obtain comprehensive human information and effectively complete human motion target recognition is a key issue in the research field of computer vision [12], [13].

CNN is a commonly used artificial neural network model [14]–[16]. Due to the large number of training data samples and the advantage of sharing weights, it can greatly reduce the complexity of neural network operations, show certain advantages in the field of machine vision and have wide application in image analysis [17]–[19]. At present, many CNN-related research results have been achieved. By using two-way long-term and short-term memory units. Literature [20] obtained video features, adopted multi-layer CNN to fuse behavior features, and got the behavior

The associate editor coordinating the review of this manuscript and approving it for publication was Dalin Zhang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

classification results through multiple convolution layers processing, and thus completed the behavior recognition. Literature [21] proposed a deep CNN for background reduction, obtained the network parameters from the training video data, constructed the optimization background model in the video, and applied it to CNN processing; by using spatial mediation filtering as the post-processing of the network output, the study constructed a better deep CNN. By using the neural network language model and the CNN shape model, Literature [22] improved the handwritten Chinese text recognition algorithm, evaluated it under the same character segmentation and classification technology, and achieved good results. Based on the topic-enhanced CNN, Literature [23] identified user interests, constructed a dual-channel CNN model to obtain information, used the Maximum Likelihood Estimation, and completed the recognition by combining with the information obtained, thereby alleviating the negative impact of noise words. After combining the residual CNN with the weighted loss function, Literature [24] analyzed with the three-dimensional model as the goal; first, the multi-view was converted into a two-dimensional view; then, the network depth was widened using the residual network; finally, the feature separation was achieved through the computation of weighted loss function, and the CNN optimization was achieved.

Based on the above existing research contents, this paper proposes a human motion target recognition algorithm based on CNN and global constraint block matching, and carries out the study of recognition by combining the CNN with the global constraint block matching approach, in order to attain more efficient recognition effects. First, the proposed algorithm is used to extract key frames of the human motion target; second, the features of key frames are fused, and the feature blocks are matched; finally, the human motion target recognition is completed by calculating the matching block score and the space weight score. The experimental results show that the proposed algorithm has superior performance, and it provides a reference for future research on computer vision. The main contributions of this paper are as follows:

(1) Combining CNN with the global constraint block matching method, human motion target recognition is completed;

(2) Key frames of human motion video are extracted, the features of key frames are obtained based on the local feature and global feature, and the feature fusion is completed;

(3) The vertical spatial information of human body is considered in an innovative manner, the matching effect of the feature block is improved, providing a basis for efficient target recognition.

(4) The experimental research is conducted by relying on UCF motion data sets, the algorithm verification analysis is finished by setting the spatial constraint weight map, key frame extraction accuracy, time of feature fusion, feature block matching accuracy, behavior recognition confusion matrix and behavior recognition F-measure, thereby enhancing the reliability of experimental results.

II. RELATED WORK

At present, many results had been produced in the research on behavior recognition. Foreign scholars Devanne *et al.* [25] proposed an approach to analyze and understand human behavior from depth video, decomposed targets using a cross-time human body posture for shape analysis, and constructed short time period of basic motions; each segment described the change in posture in terms of human motion and the appearance of the depth around the knuckles of the hand; finally, the Naive Bayesian Classifier was used to capture the dynamic characteristics of basic motion features and to change human behaviors; the method was proved to have good overall performance, yet with poor feature matching effect. Coupeté *et al.* [26] proposed an online gesture recognition technology that enabled human-robot collaboration in an industrial context; the effectiveness of the method had been revealed in several documents, but it has poor feature fusion effect. Ding *et al.* [27] proposed a linear dynamic system of three-dimensional skeleton motion recognition for modeling time series tensor observations, estimated the parameters of linear dynamic system as a motion description using Tucker decomposition, and did an in-depth analysis of motion recognition; however, the target motion recognition rate still needs to be improved. Chen *et al.* [28] proposed a method of CNN acceleration using hardware/software co-design technology, and created an embedded system using the ARM processor and FPGA structure, so as to balance diversity and flexibility; the method has better overall effect especially for the system research with limited resources, but the accuracy rate of target recognition needs to be further improved.

At home, many results had also been produced in the research on CNN and behavior recognition. Liangliang *et al.* [29] proposed an improved video representation method of feature graph strings to obtain ordered behavior video feature strings, matched human behavior and completed recognition based on the graph matching method and dynamic time warping method, proving that the method had low recognition accuracy. Lifei *et al.* [30] combined double flow network and CNN for behavior recognition, obtained the spatial-temporal features of the video using the multi-scale input 3D convolution fusion network, fused the data of different scales, and completed the behavior recognition by building the double flow model, proving that the algorithm had poor data fusion effect. Guisheng *et al.* [31] extracted the structural information and inherent features of the sketch using the convolution kernel, increased the depth of the network, and enhanced the data diversity, proving that the method could improve the recognition accuracy. Guoliang and Xiaoxiang [32] proposed a behavior recognition method based on mobile phone's multi-source sensor, extracted the current behavior feature using the sensor, obtained the feature vector, and verified the performance of the algorithm based on different feature matching behaviors, proving that the algorithm showed good recognition effect over persistent motions yet its accuracy rate should be further improved.

In order to address the problems of above research methods, the CNN was combined with global constraint block matching to complete human motion target recognition, and the global constraint block is matched with the feature block from the perspective of spatial information in the vertical direction of the human body, thereby improving the matching effect and providing a good basis for subsequent target recognition.

III. HUMAN MOTION TARGET RECOGNITION ALGORITHM BASED ON CNN AND GLOBAL CONSTRAINT BLOCK MATCHING

A. EXTRACTION OF KEY FRAMES FROM THE HUMAN MOTION VIDEO

Human motion is characterized by certain continuity, leading to a large number of redundant frames in motion video [33]. In order to quickly obtain valid information, the motion abrupt frame will be used as a key frame for extracting key information from the human video [34], [35].

The human motion video is described as s, the first frame is set as a reference frame s_0, and the other frames are compared with the reference frame by taking the motion angle and position as indicators. If there is a large difference, the frame is regarded as a key frame, otherwise it is deemed as a redundant frame. The specific expression is as follows:

$$\left| d_{s_i} - d_{s_0} \right| > \text{thd\_}pos \tag{1}$$

$$\left| \varphi_{s_i} - \varphi_{s_0} \right| > \text{thd\_}ang \tag{2}$$

where, d_{s_0} and d_{s_i} represent the reference frame position and the i-th frame video human motion position respectively. φ_{s_0} and φ_{s_i} represent the reference frame angle and the i-th frame video human motion angle respectively. thd_*pos* is the moving target position difference threshold. thd_*ang* is the threshold value of moving target angle difference.

If any of the above formulas (1) and (2) satisfy one of the conditions, that is, the distance between the current frame and the reference frame is greater than the threshold of the position difference, or the motion offset angle of the current frame and the reference frame is greater than the threshold of angle difference, then the frame is deemed as a key frame [36], [37].

B. CNN-BASED KEY FRAME FEATURE FUSION

CNN is a typical deep learning neural network [38], [39]. It consists of the convolution layer, pooling layer, normalized layer and Concat layer. The convolution layer extracts image features by using the convolution kernel, which has more neurons than the shallow neural network. In addition, CNN can extract data and the extraction process can be designed independently [40], [41], so this approach shows certain advantages in data analysis. The pooling layer completes data downsampling and reduces the dimensions; the normalized layer establishes a competition mechanism through the local neurons, and improves the generalization ability of the network; the Concat layer can avoid data over-fitting problems [42], [43].

Usually, there are highly complex data features in the last layer of CNN [44], [45]. In order to obtain better results with generalization capabilities and comprehensively obtain the key frame features, this study independently designs a method of extracting the local feature and the global feature by using the convolution layer of CNN, and then inputs the feature extraction results into CNN for fusion processing, as follows:

1) LOCAL FEATURE

Local feature is a description of human appearance, partial occlusion and visual changes, and it can be reflected by interest point detection.

Suppose that $P = \{p_1, p_2, \ldots, p_n\}$ is the corresponding spatial-temporal interest point at any time in the sequence $g(x, y)$ of the human motion key frame, and the total number is n, then the interest point can be represented by $p_i(x_i, y_i)$.

The center point of human motion in the key frame is set to the pole L, and the interest point $p_i(x_i, y_i)$ can be converted into the following coordinate form in the coordinate system of the pole L, which is expressed by the following formula:

$$\begin{cases} r_i = \sqrt{(y_i - y_o)^2 + (x_i - x_o)^2} \\ \alpha_i = \arctan\left(\dfrac{y_i - y_o}{x_i - x_o}\right) \end{cases} \quad i = 1, 2, \ldots, n \tag{3}$$

where, r_i represents the translation distance of the interest point $p_i(x_i, y_i)$, α_i is the angle formed by the pole L and the interest point $p_i(x_i, y_i)$, and (x_i, y_i) and (x_o, y_o) are the coordinates of interest point and the center point of human body, respectively.

2) GLOBAL FEATURE

In the global feature extraction, this paper starts with the contour line of human body and uses the curvature function for description. Besides, the study describes the human contour line in the key frame sequence, and randomly samples M points from the contour line. The coordinate form of these sampling points is denoted as (a_i, b_i).

A sampling point $c(a_1, b_2)$ is selected. The curvature is defined as the change rate of tangential angle of the sampling point relative to the arc length, and the curvature function is expressed as:

$$q(c) = \frac{d}{dc}\theta(c) \tag{4}$$

where, $\theta(c)$ is the tangential angle.

Fourier transform coefficient F is used to describe the curve function. Since Fourier transform coefficient is symmetric, only the coordinate system of positive frequency needs to be considered, thereby improving the computation efficiency [46], [47]. The global feature based on the curvature function is described as:

$$G = \{q(c) \cdot |S_1| \, |S_2| \ldots |S_M|\} \tag{5}$$

where, S_i represents the curvature function.

3) CNN FEATURE FUSION PROCESSING

In this study, the Concat layer is added to the CNN, the last three fully connected layers are deleted, and the extracted local and global key frame features are input into it. As a result, the key frame feature fusion processing is performed based on the CNN designed above [48], [49].

In this study, 20 frames of training video images are randomly selected, the video image set is represented as $(\lambda_o, \lambda_1, \ldots, \lambda_{i-1})$, and the size of single frame image in the video sequence is designed as 225 × 225. At this time, the input data scale in CNN is 20 × 225 × 225 × 3 (20 is the video image sequence length and 3 is the channel), and the CNN model parameters are trained under this condition, and a total of 6 convolution layers are designed. Among them, the step length of the first convolution layer is 2; the step length of the second convolution layer is 2; the step lengths of the third, fourth and fifth convolution layers are 1; the step length of the sixth convolution layer is 1. In order to ensure the non-linearity of the model, an activation function is added after each convolution layer; in order to speed up the training speed of the neural network, a normalized layer is also added after each pooling layer [50], [51].

The fifth convolution layer represents the deep feature map. In this study, this shallow feature map of the third convolution layer is fused with that of the fifth convolution layer [52], [53]. This fusion process is implemented based on the Concat layer. The feature fusion calculation formula is as follows:

$$f_i = H\left([\lambda_0, \lambda_1, \ldots, \lambda_{i-1}]\right) \tag{6}$$

where, $H(\cdot)$ is nonlinear transform function, and f_i represents the results of key frame feature fusion.

C. TARGET RECOGNITION BASED ON GLOBAL CONSTRAINT MATCHING BLOCK

1) MATCHING OF HUMAN MOTION TARGET FEATURE BLOCK

Human motion target recognition needs to be completed by matching different key frame feature blocks. This study extracts the matching blocks of two different images, calculates the distance between the two matching blocks, and obtains the closest matching block to complete the human motion target feature block matching.

Subsequently, the features obtained after the key frame fusion are combined into two different feature block sets $F\{F_1, F_2, \ldots, F_n\}$ and $F'\left\{F'_1, F'_2, \ldots, F'_n\right\}$, where, $F_n = \{f_1, f_2, \ldots, f_n\}$, $F'_n = \left\{f'_1, f'_2, \ldots, f'_n\right\}$.

A set of human motion image key frame feature sequences are selected in the feature block set $F\{F_1, F_2, \ldots, F_n\}$, and a rectangular coordinate system is set; the l-th center coordinate of the image feature block is set to (u, v) and the same vertical coordinate (u, v); under different conditions of the abscissa u (the maximum value of u is U), the feature block $F'(l, u)$ of the l-th image in the feature block set $F'\left\{F'_1, F'_2, \ldots, F'_n\right\}$ is represented as:

$$F'(l, u) = \left\{g^l_{u,v}\right\}, \quad u = (1, 2, \ldots, U) \tag{7}$$

where, $\left\{g^l_{u,v}\right\}$ represents the l-th image feature point set.

In order to avoid the impact of spatial changes in video image sequences on human motion target recognition, this study, based on the abscissa of the search space, allows a certain fluctuation direction of the ordinate and expands the search range within a reasonable range. The search area can be defined as:

$$Area(u) = \left\{g^l_{u,v} | v = v-1, \ldots, v \ldots, v+1\right\} \tag{8}$$

where, The value of v ranges from $v + 1 < V$ and $v - 1 > 0$, and V represents the maximum vertical coordinate of the image feature block.

The l-th image feature block is used as the feature block $F'(l, u)$ to be matched in the feature block set $F'\left\{F'_1, F'_2, \ldots, F'_n\right\}$, and the closest matching feature block $\mathrm{M}_{F'(l,u)}$ obtained by $F'(l, u)$ search is:

$$\mathrm{M}_{F'(l,u)} = \arg\min \left\{g | g = d\left(g^l_{u,v}\right), g^l_{u,v} \in Area(u)\right\} \tag{9}$$

where, g represents the distance between feature blocks. $d\left(g^l_{u,v}\right)$ represents the distance between the center point of the l-th image feature block and the center point of the adjacent feature block.

2) CALCULATION OF SPATIAL CONSTRAINT WEIGHT

From the perspective of mathematics, constraint is one of the effective methods to solve the optimal solution. After completing the matching of human motion target feature blocks, this paper uses the definition of spatial constraint to limit the scope of target recognition, weaken invalid matches, and further increase the strong matching relationship, which lays the foundation for improving the accuracy of human motion target recognition. Spatial constraint means that a feature block selects the nearest feature block in the same row of the image, and a distance vector is formed; the feature block of another image selects the nearest feature block in the same way and a distance vector is formed; the difference between two distance vectors is calculated according to Euclidean distance, and the spatial constraint weight can be obtained based on the results.

Since human motion is usually upright and the vertical spatial data is closer for the same human motion target, this paper calculates the spatial constraint weight of the human motion feature block from the vertical direction, thereby compensating for the poor matching effect of feature block and make the calculation result more comprehensive. The calculation process is as follows:

Assuming that the feature block in the feature block set $F'\left\{F'_1, F'_2, \ldots, F'_n\right\}$ of the l-th image is $F'(l, u)$, this study selects the closest feature block $F'(k, u)$ in the same row in its own image, and gives the Euclidean distance

$d\left(F'(l,u), F'(k,u)\right)$ of the two image feature blocks. Based on the Euclidean distance, the distance $d_{l,k}$ of the two image feature blocks can be calculated:

$$d_{l,k} = \min \left\{ \begin{array}{l} d\left(F'(l,u), F'(1,u)\right), \\ d\left(F'(l,u), F'(2,u)\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots, \\ d\left(F'(l,u), F'(k,u)\right) \end{array} \right\} \quad (10)$$

The distance between the feature blocks of the two images calculated according to Formula (10), which represents the closest distance required by the feature block $F'(l,u)$ to search for similar feature blocks. In order to directly and accurately represent the similarity of the two feature blocks in the spatial structure, a quantitative numerical method is used to represent them, and the spatial constraint weight $w\left(F'\right)$ of matching feature block was calculated:

$$w\left(F'\right) = \sum_{k=1}^{Q}\left(\left|F'(l,u) - F'(k,u)\right|\right) \quad (11)$$

where, Q represents the foreground area of motion target recognition. The smaller the spatial constraint weight $w\left(F'\right)$, the higher the similarity between the two feature blocks in the vertical direction and the higher the matching rate.

3) TARGET RECOGNITION

The key to human motion target recognition research lies in the matching of two images. This matching problem is represented by a matching score, which can get more accurate results. In this study, human motion target recognition is mainly calculated from the matching block score and the spatial constraint weight score, and then the target recognition is completed according to the cumulative result of the score. The specific steps are as follows:

Input: According to the feature fusion results calculated through CNN, enter image feature block matching data and weight calculation result.

Output: The result of human motion target recognition is represented by the matching score.

The motion target data information is initialized, the human motion target is identified and studied, and it is described with the following algorithm:

1) In human motion target recognition, due to environmental factors such as background and lighting, the inaccurate target matching issue is prone to occur, that is, inaccurate distance calculation occurs in the matching block search. To avoid this problem, the distance problem is converted into a scoring problem for calculation. Besides, Gaussian functions need to be used in the conversion process; Calculated as follows:

$$G(l,k) = \exp\left(-\frac{d\left(F'(l,u), F'(k,u)\right)}{\chi^2}\right) \quad (12)$$

where, χ is the bandwidth of the Gaussian function.

2) The score of the feature block $F'(l,u)$ for its matching block $F'(k,u)$ is expressed as:

$$score\left(F'(l,u), F'(k,u)\right) = G(l,k) \quad (13)$$

3) According to the spatial constraint weight calculation method in Section III(C (2)), the spatial weights of two different images are calculated, and the weight calculation results are converted into a numerical form $w_1\left(F'\right)$ suitable for being expressed by score. Such as formula (14).

$$w_1\left(F'\right) = \exp\left(-\frac{w\left(F'\right)^2}{2\chi^2}\right) \quad (14)$$

4) The score of spatial constraint weight can be obtained according to the last step. Such as formula (15).

$$score'\left(F'(l,u), F'(k,u)\right) = \sum_{k=1}^{Q}\left(score\left(F'(l,u), F'(k,u)\right), w_1\left(F'\right)\right) \quad (15)$$

5) The final cumulative score can be obtained by combining the score of matching block and the score of spatial constraint weight:

$$S_{total} = score\left(F'(l,u), F'(k,u)\right) + score'\left(F'(l,u), F'(k,u)\right) \quad (16)$$

The higher the cumulative score S_{total}, the more likely two images are to be the same target

6) End.

According to the above steps, human motion target recognition based on the CNN and the global constraint block matching can be completed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effectiveness of the research algorithm in this paper, an experimental analysis is required. The specific arrangements are as follows.

A. EXPERIMENTAL ENVIRONMENT AND DATASET

In this experiment, Windows 10 operating system is used, and the Opencv open source computer vision database is adopted for software, which are more convenient to complete motion recognition. In addition, the experiment is carried out in a GPU accelerated environment by using a deep learning database with a python interface.

The experimental data is selected from the UCF (https://www.crcv.ucf.edu/data/UCF101.php) motion data set. This data set contains a large number of videos, diverse scenes and a variety of motions, such as kicking, diving, running, walking, and archery; it also contains rich human motion data, and most data in this data set are taken from real scenes. Therefore, this data set has great practical significance for experimental analysis. The experimental data parameter settings are shown in Table 1.

TABLE 1. Experimental data parameter settings.

Parameter	Value
Number of action videos	2000 万
Image pixels	227*227
Video frame	400 frame
Frame rate	25frame/s
Scenes	10 piece

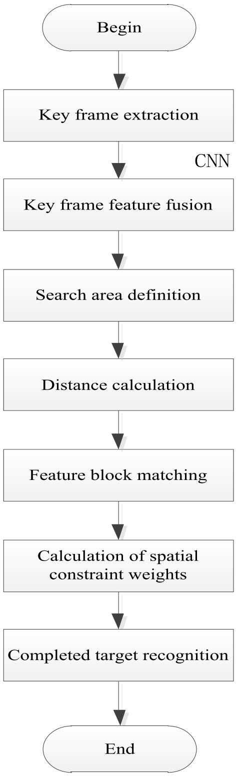


FIGURE 1. Human motion target recognition process.

A total of 20 million human motion videos are selected in the experiment, of which 10 million videos are used as the test set and 10 million videos are used as the training set. Before testing the experimental data, this study first extracts the optical flow and RGB frames of video data, processes them in advance, clips the video data into short videos at a frame rate of 25 frames per second, and adjusts the pixels to 227*227, thereby ensuring that the human motion image features are not lost.

B. EXPERIMENTAL INDICATORS

The following experimental indicators are selected to verify the algorithm in this paper, as follows:

1) Spatial constraint weight map: This study selects the vertical direction and uses the definition of spatial constraint to further enhance the matching degree of human motion target, providing a basis for target recognition. This is the key content of the algorithm research in this paper. In order to verify the effectiveness of the algorithm, the walking human targets are selected from the database as objects, and the spatial constraint weight map is designed to compare the performance of the algorithm;
2) Accuracy of key frame extraction: Key frame extraction is the first step of the algorithm calculation in this paper, which can effectively avoid the interference of useless information. The key frame extraction accuracy is used as an indicator. The proposed algorithm is then compared with the algorithms [12], [13], [15], [16] and [17], and the calculation formula for the key frame extraction accuracy is as follows:

$$D_{accu} = \frac{t_i}{t} \times 100\% \tag{17}$$

where, t represents the actual number of key frames, and t_i represents the number of key frames extracted according to the proposed algorithm.
3) Feature fusion time consumption: This paper uses the CNN to complete the key frame feature fusion, which lays the foundation for feature block matching. According to Formula (6), this study obtains the feature fusion result, calculates the feature fusion time, compares the proposed algorithm with algorithms in literature [12], [13], [15], [18] and [19], so as to verify the performance of the proposed algorithm.
4) Feature block matching accuracy: Before the recognition of human motion target is completed, feature block matching is required, so feature block matching accuracy is a direct factor that determines the target recognition effect. In this experiment, feature block matching accuracy is used as an indicator to verify the performance of the proposed algorithm. The formula for feature block matching accuracy is as follows:

$$\text{Accuracy} = \frac{M_{mate}}{M_{act}} \times 100\% \tag{18}$$

where, M_{mate} is the value of the algorithm matching result, and M_{act} is the value of the actual matching result.
5) Behavior recognition confusion matrix: The confusion matrix is a standard format for accuracy evaluation. In this experiment, a behavior recognition confusion matrix is established, and the correct recognition rate is calculated through the confusion matrix analysis;
6) F-measure for behavior recognition: F-measure is a statistic amount and it is a weighted harmonic mean of recall rate and accuracy. In order to verify the accuracy of the recognition results in this paper, the F-measure of the proposed algorithm is compared with that of the algorithms in literature [10], [11], [15], [18] and [19].

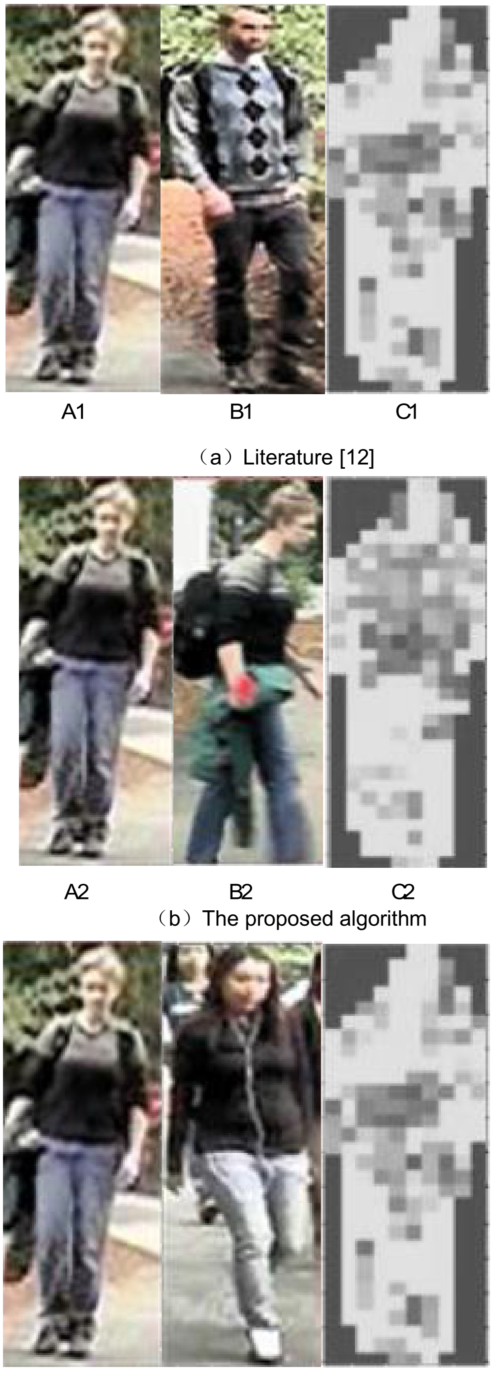
$$\text{F-measure} = \frac{R \cdot Accu}{R + Accu} \tag{19}$$

where, R represents the recall rate of behavior recognition, and $Accu$ represents the accuracy of behavior recognition.

C. EXPERIMENTAL RESULTS

1) SPATIAL CONSTRAINT WEIGHT MAP

A pedestrian walking image is selected from the database as the original image. In this study, a spatial constraint weight map is established, and the proposed algorithm is compared with Literature [12] and Literature [19]. It is shown in Figure 2.



（a）Literature [12]

（b）The proposed algorithm

（c）Literature [19]

FIGURE 2. Spatial constraint weight map.

Figure 2 (a) is a group of images composed of three images, which are defined as A1, B1, and C1 from left to right. Similarly, the three images in Figure 2 (b) are defined as A2, B2, and C2 in this order, the three images in Figure 2 (c) are defined as A3, B3, and C3 in this order.

In Figures 2 (a), (b), and (c), the images A1, A2, and A3 are the same, and they are all original images.B1, B2, B3 represent pedestrian images similar to the original image, C1, C2, and C3 are an evaluation image of the similarity of the two pedestrians. Among them, the more and denser the square spots, the smaller the spatial constraint weight, and the greater the similarity between the two pedestrians. According to Figure 2, most of the pedestrians obtained from the matching of different algorithms are close to the features in the original image, but the behaviors obtained from the matching of proposed algorithm are consistent with the pedestrians in the original image; as a result, the pedestrians obtained by matching through the algorithms in Literature [12] and Literature [19] are still quite different from those in the original image. This is because this paper considers the differences in the vertical pedestrian structure, which makes up for the poor feature block matching and obtains good results.

2) KEY FRAME EXTRACTION ACCURACY RATE

The proposed algorithm is compared with the algorithms in Literature [12], [13], [15], [16] and [17] in terms of key frame extraction accuracy rate.It is shown in Figure 3.

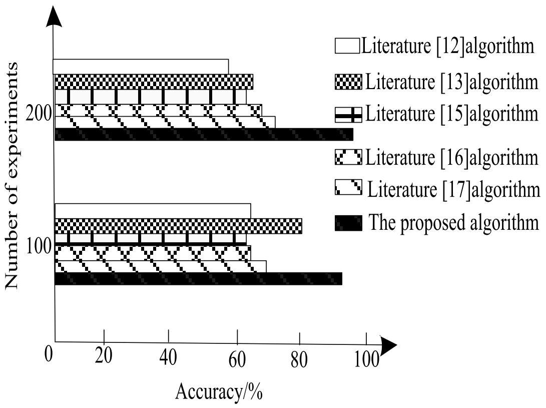


FIGURE 3. Comparison of key frame extraction accuracy.

According to Figure 3. During the experiment, the key frame extraction accuracy rate of the proposed algorithm is always above 90%. In comparison, the algorithm in Literature [12] has an extraction accuracy rate of about 60%; the algorithm in Literature [13] has a high extraction accuracy rate of 60%-80%, the algorithm in Literature [15] and Literature [16] accuracy rate are about 60%, the algorithm in Literature [17] has the maximum extraction accuracy rate of 70%. According to the data description, the proposed algorithm has a high key frame extraction accuracy rate, which indicates that the motion target position difference threshold and angle difference threshold of the proposed algorithm are selected in an appropriate manner, and they can accurately complete the key frame feature extraction.

3) FEATURE FUSION TIME CONSUMPTION

Time consumption is an important way to measure the quality of the fusion result. In this study, the feature fusion time consumption of the proposed algorithm is compared with the algorithms of other literature. It is shown in Figure 4.

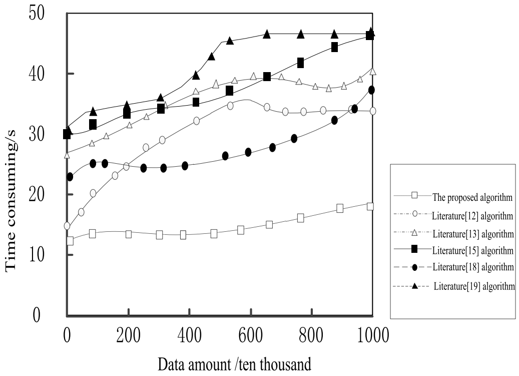


FIGURE 4. Time-consuming comparison of feature fusion.

With the increase of the data amount, the key frame feature fusion time consumption of the algorithm has shown an overall upward trend. Among them, the time consumption of the algorithm in Literature [19] rises significantly and reaches the highest; the feature fusion time consumption of the algorithm in Literature [19] is 47s when the data amount is 10 million; followed by the algorithm in Literature [15], the maximum time consumption also reaches 47s; the time consumption of the algorithm in Literature [12] rises significantly before the data amount is 60 million and reaches a steady state after the data amount is 60 million, with the maximum time consumption of around 32s; the time consumption curves in Literature [13] and Literature [18] are relatively flat, with the maximum time consumption of 12-15s. according to the figure, the proposed algorithm has a smooth time consumption curve and time consumption is 12-15s, indicating that the proposed algorithm is obviously better than other algorithms in the above literature. This is because CNN is used in this study for feature fusion, and the CNN model parameters are trained to help the algorithm quickly complete key frame feature fusion.

4) FEATURE BLOCK MATCHING ACCURACY

Feature block matching is a prerequisite for human motion target recognition. To verify the performance of the proposed algorithm, this paper compares the proposed algorithm with the algorithms in Literature [13], [19] and [20] in terms of feature block matching accuracy. It is shown in Figure 5.

Figure 5 can be seen that the average value of the feature block matching accuracy of the proposed algorithm is more than 80%, while the feature block matching accuracy of the algorithms in literature [13], [19], and [20] is low with an average value of about 60%, so the proposed algorithm shows obvious advantages. This is because after matching the feature block, this paper calculates the spatial constraint weight and calculates the weights from the vertical direction, which improves the matching accuracy.

5) BEHAVIOR RECOGNITION CONFUSION MATRIX

The confusion matrix, called the error matrix, is represented by n rows and n columns. In this study, each measured value is compared with the predicted value. The sum of each row represents the true number of samples in that category, and the sum of each column represents the number of samples identified as that category. Besides, this experiment uses the behavior recognition confusion matrix to represent the behavior recognition rate of the algorithm. It is shown in Table 2.

TABLE 2. Behavior recognition confusion matrix.

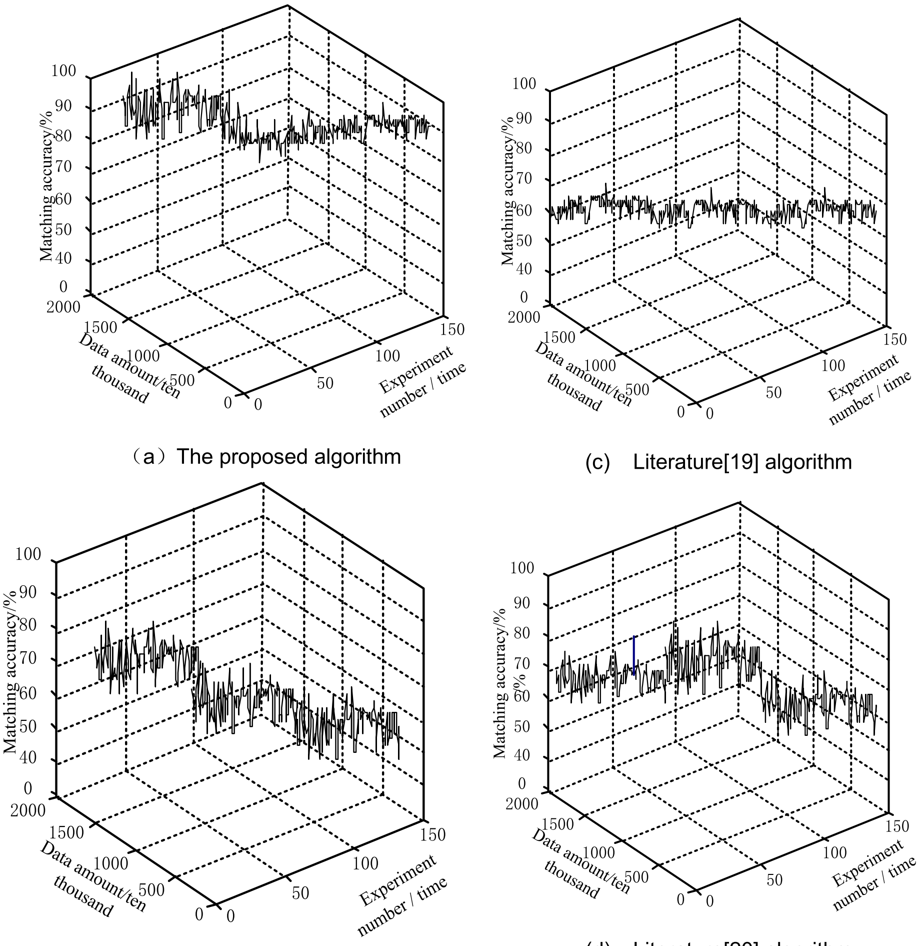
		Algorithm identification value				
		Kicking	Diving	Running	Walking	Archery
Actual value	Kicking	0.95	0.00	0.05	0.00	0.00
	Diving	0.05	0.90	0.00	0.05	0.00
	Running	0.00	0.05	0.85	0.00	0.10
	Walking	0.00	0.00	0.00	1.00	0.00
	Archery	0.05	0.00	0.00	0.05	0.90

According to the behavior recognition confusion matrix in Table 2, this study visualizes the human motion target recognition results, and uses diagonal elements to represent the correct recognition rate. It can be seen from Table 2 that for five motions of kicking, diving, running, walking and archery, the correct recognition rates of the proposed algorithm are 0.95, 0.90, 0.85, 1.00 and 0.90, respectively, indicating that the proposed algorithm has very good human motion target recognition capabilities.

6) F-MEASURE OF BEHAVIOR RECOGNITION

F-measure for behavior recognition of the proposed algorithm is calculated and compared with the F-measure calculated by the algorithms in literature [10], [11], [15], [18] and [19].

According to the F-measure comparison results of behavior recognition in Table 3, under the condition that the video frame takes different values, the proposed algorithm always has the largest F-measure of behavior recognition, with an average value of about 0.95. In contrast, the average value of F-measure calculated according to the algorithm in Literature [10] is about 0.78; the value calculated according to the algorithm in Literature [11] is about 0.72; the value calculated according to the algorithm in Literature [15] is about 0.77; the value calculated according to the algorithm in Literature [18] is about 0.82; the value calculated according to the algorithm in Literature [19] is about 0.86. According to



（a）The proposed algorithm

(c) Literature[19] algorithm

(b) Literature[13] algorithm

(d) Literature[20] algorithm

FIGURE 5. Comparison of feature block matching accuracy.

the data analysis, the proposed algorithm shows significant advantages in F-measure and has certain advantages in human motion target recognition.

V. DISCUSSION

In order to further promote the development of computer vision research, to more fully understand the motion video information, and to obtain a more effective human body target recognition method, this paper proposes a human body target recognition algorithm based on convolutional neural network and global constraint block matching. The experimental results show that the proposed algorithm has a good fusion of human features, the target matching result is consistent with the original image, the recognition accuracy is high, and it can effectively avoid problems such as the number of images, camera angle, and light.

Human motion recognition has attracted attention due to its extensive application in human-computer interaction, visual surveillance, and video indexing. Therefore, the study of human motion target recognition in this paper is of great significance, and some related literature has achieved some results.

TABLE 3. Comparison of F-measure of behavior recognition.

Algorithm	Video frame/ frame			
	100	200	300	400
literature [10]algorithm	0.76	0.71	0.81	0.82
literature [11]algorithm	0.66	0.72	0.75	0.76
literature [15]algorithm	0.75	0.80	0.74	0.80
literature [18]algorithm	0.72	0.79	0.87	0.89
literature [19]algorithm	0.80	0.86	0.86	0.90
The proposed algorithm	0.93	0.92	0.95	0.98

Abdelbaky and Aly [54] proposed a new method for human motion recognition based on principal component analysis network, which uses the motion energy template to appropriately represent the time information of the input video, and calculates a plurality of short-term motion energy image templates to capture human motion information. Yanhu *et al.* [55] described the visual behavior recognition deeply based on the evolution of database. Han *et al.* [56] used human-computer interaction to study human behavior and human intention recognition.

These literatures have achieved certain results, but compared with the proposed algorithm, it has obvious advantages in key frame extraction, feature fusion, feature block matching, human behavior recognition, etc., which can provide data basis for human behavior research and help further computer research and development.

However, in the research of the recognition algorithm in this paper, the influencing factors such as human motion posture were not considered, and the comprehensiveness of the research results is insufficient. In the next research, we need to comprehensively consider more influencing factors of human motion recognition, obtain more comprehensive research results, and provide a reference basis for computer vision research.

VI. CONCLUSION

Human motion target recognition is one of the key research topics in the field of computer vision. In order to solve the shortcomings of traditional target recognition algorithms, this paper proposes a human motion target recognition algorithm based on the CNN and the global constraint block matching, extracts motion video key frame, fuses the key frame features based on the CNN, completes feature block matching according to the definition of spatial constraint, thereby realizing human motion target recognition. 10 million data were selected from the UCF motion data sets as a test set to verify the proposed algorithm process and research results. The key frame extraction accuracy of the proposed algorithm is as high as 98%, and the feature fusion takes less than 15s, which is much lower than other literature algorithms.The experimental results show that the proposed algorithm takes into account the differences in pedestrian structure in the vertical direction and makes up for the lack of feature block matching. The matching image is consistent with the original image, and the average value of the feature block matching accuracy is more than 80%. The correct recognition rate of human moving targets is not less than 0.85, and the overall recognition effect is better.

REFERENCES

[1] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.

[2] G. Ge, Y. Cai, and Q. Dong, "A flexible pressure sensor based on rGO/polyaniline wrapped sponge with tunable sensitivity for human motion detection," *Nanoscale*, vol. 10, no. 21, pp. 10033–10040, 2018.

[3] H.-J. Byun and S. G. Shon, "Implementation about thread and Internet-based motion receiving imitation controller for humanoid," *Int. J. Secur. Appl.*, vol. 10, no. 1, pp. 65–74, Jan. 2016.

[4] W. Ke, W. Huiqin, S. Yue, M. Li, and Q. Fengyan, "Banknote image defect recognition method based on convolution neural network," *Int. J. Secur. Appl.*, vol. 10, no. 6, pp. 269–280, Jun. 2016.

[5] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 15, 2019, doi: 10.1109/TCSVT.2019.2947482.

[6] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 30, 2019, doi: 10.1109/TPAMI.2019.2932058.

[7] J. Won and J. Lee, "Shadow theatre: Discovering human motion from a sequence of silhouettes," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, 2016.

[8] J. Zhu, Y. Shang, and Y. Xia, "An atypical MAGUK GK target recognition mode revealed by the interaction between DLG and KIF13B," *Structure*, vol. 24, no. 11, pp. 1876–1885,2016.

[9] Y. Yin, J. Xia, Y. Li, Y. Xu, W. Xu, and L. Yu, "Group-wise itinerary planning in temporary mobile social network," *IEEE Access*, vol. 7, pp. 83682–83693, 2019.

[10] J. Li, X. Zhang, Z. Wang, X. Chen, J. Chen, Y. Li, and A. Zhang, "Dual-band eight-antenna array design for MIMO applications in 5G mobile terminals," *IEEE Access*, vol. 7, pp. 71636–71644, 2019.

[11] L. Guido and T. Anna, "A structural view of microRNA-target recognition," *Nucleic Acids Res.*, vol. 44, no. 9, pp. 1–8, 2016.

[12] Y. Guo, H. Xiao, and Q. Fu, "Least square support vector data description for HRRP-based radar target recognition," *Appl. Intell.*, vol. 46, no. 2, pp. 1–8, 2016.

[13] T. Hachaj, M. R. Ogiela, and K. Koptyra, "Human actions recognition from motion capture recordings using signal resampling and pattern recognition methods," *Ann. Oper. Res.*, vol. 265, no. 2, pp. 1–17, 2018.

[14] B. Zhu, Y. Chen, Y. Cai, H. Tian, and T. Wang, "Network security situation prediction system based on neural network and big data," *Int. J. Secur. Appl.*, vol. 11, no. 1, pp. 93–108, Jan. 2017.

[15] S. Bal, M. K. Rai, H.-J. Kim, and R. Saha, "An adaptive parallel interference mitigation technique using artificial neural network in complementary coded MC-CDMA," *Int. J. Future Gener. Commun. Netw.*, vol. 11, no. 4, pp. 69–78, Jul. 2018.

[16] Y. Cho and S. Woo, "Automated ROI detection in left hand X-ray images using CNN and RNN," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 7, pp. 81–92, Jul. 2018.

[17] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, "Context-aware QoS prediction with neural collaborative filtering for Internet-of-Things services," *IEEE Internet Things J.*, early access, Dec. 2, 2019, doi: 10.1109/JIOT.2019.2956827.

[18] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 661–674, Feb. 2020, doi: 10.1109/TNNLS.2019.2908982.

[19] Y. Jin, X. Guo, Y. Li, J. Xing, and H. Tian, "Towards stabilizing facial landmark detection and tracking via hierarchical filtering: A new method," *J. Franklin Inst.*, Jan. 2020, doi: 10.1016/j.jfranklin.2019.12.043.

[20] G. Rui, W. Zhaohui, and X. Xin, "Behavior recognition based on multi-layer convolutional neural network features and bidirectional long and short-term memory units," *Control theory Appl.*, vol. 34, no. 6, pp. 790–796,2017.

[21] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.

[22] Y.-C. Wu, F. Yin, and C.-L. Liu, "Improving handwritten chinese text recognition using neural network language models and convolutional neural network shape models," *Pattern Recognit.*, vol. 65, pp. 251–264, May 2017.

[23] D. Yumeng, Z. Weinan, and L. Ting, "User interest recognition based on topic enhanced convolutional neural network," *Comput. Res. Develop.*, vol. 55, no. 1, pp. 188–197,2018.

[24] L. Y. Shengyan, P. Xiang, and L. Fuchang, "Optimization of residual convolution network for feature extraction of 3D model views," *J. Comput. Aided Des. Graph.*, vol. 31, no. 6, pp. 936–942, 2019.

[25] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, and A. D. Bimbo, "Motion segment decomposition of RGB-D sequences for human behavior understanding," *Pattern Recognit.*, vol. 61, pp. 222–233, Jan. 2017.

[26] E. Coupeté, F. Moutarde, and S. Manitsaris, "Multi-users online recognition of technical gestures for natural human–robot collaboration in manufacturing," *Auto. Robots*, vol. 43, no. 6, pp. 1309–1325, Aug. 2019.

[27] W. Ding, K. Liu, E. Belyaev, and F. Cheng, ''Tensor-based linear dynamical systems for action recognition from 3D skeletons,'' *Pattern Recognit.*, vol. 77, pp. 75–86, May 2018.

[28] A. T.-Y. Chen, M. Biglari-Abhari, and K. I.-K. Wang, ''Convolutional neural network acceleration with hardware/software co-design,'' *Appl. Intell.*, vol. 48, no. 3, pp. 1288–1301,2017.

[29] S. Liangliang, L. Dong, and T. Jun, ''Recognition of human behavior based on improved feature map string method,'' *Eng. Sci. Technol.*, vol. 48, no. 6, pp. 165–171, 2016.

[30] S. Lifei, W. Liguo, and W. Lingfeng, ''Behavior recognition method of multi-scale input 3D convolution fusion dual flow model,'' *J. Comput. Aided Des. Graph.*, vol. 30, no. 11, pp. 99–108, 2018.

[31] Y. Guisheng, Y. Xue, and W. Yuhua, ''Handwritten sketch recognition based on convolutional neural network,'' *J. Jilin Univ. Inf. Sci. Ed.*, vol. 37, no. 4, pp. 417–425, 2019.

[32] C. Guoliang and C. Xiaoxiang, ''Indoor fire pedestrian behavior recognition based on mobile multi-source sensors,'' *J. Tongji Univ.*, vol. 47, no. 3, pp. 414–420, 2018.

[33] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, ''QoS prediction for service recommendation with deep feature learning in edge computing environment,'' *Mobile Netw. Appl.*, Apr. 2019, doi: 10.1007/s11036-019-01241-7.

[34] W. S. Yoo and B. H. Lee, ''Dynamic key-frame selection technique for enhanced odometry estimation based on laser scan similarity comparison,'' *Electron. Lett.*, vol. 53, no. 13, pp. 852–854, Jun. 2017.

[35] R. S. Rokade and D. D. Doye, ''Spelled sign word recognition using key frame,'' *IET Image Process.*, vol. 9, no. 5, pp. 381–388, May 2015.

[36] M. W. Haskell, S. F. Cauley, B. Bilgic, J. Hossbach, D. N. Splitthoff, J. Pfeuffer, K. Setsompop, and L. L. Wald, ''Network accelerated motion estimation and reduction (NAMER): Convolutional neural network guided retrospective motion correction using a separable motion model,'' *Magn. Reson. Med.*, vol. 82, no. 4, pp. 1452–1461, Oct. 2019.

[37] Y. Xu, D. Li, Z. Wang, Q. Guo, and W. Xiang, ''A deep learning method based on convolutional neural network for automatic modulation classification of wireless signals,'' *Wireless Netw.*, vol. 25, no. 7, pp. 3735–3746, Oct. 2019.

[38] X. Xie and A. Wang, ''Study of the grid resource neural network reservation technology based on SLA,'' *Int. J. Grid Distrib. Comput.*, vol. 10, no. 1, pp. 21–30, Jan. 2017.

[39] Y. Son, S. Oh, and Y. Lee, ''Hybrid deep neural network based performance estimation method for efficient offloading on IoT-cloud environments,'' *Int. J. Grid Distrib. Comput.*, vol. 11, no. 7, pp. 23–30, Jul. 2018.

[40] A. Wang, Y. Wang, and Y. Chen, ''Hyperspectral image classification based on convolutional neural network and random forest,'' *Remote Sens. Lett.*, vol. 10, no. 11, pp. 1086–1094, Nov. 2019.

[41] J. Ishioka, Y. Matsuoka, S. Uehara, Y. Yasuda, T. Kijima, S. Yoshida, M. Yokoyama, K. Saito, K. Kihara, N. Numao, T. Kimura, K. Kudo, I. Kumazawa, and Y. Fujii, ''Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm,'' *BJU Int.*, vol. 122, no. 3, pp. 411–417, Sep. 2018.

[42] Z. Qsuanhua, L. Xiaoli, and Z. Xuemei, ''Fuzzy clustering image segmentation based on spatial constraint student's-t hybrid model,'' *Control Decis.*, vol. 31, no. 11, pp. 2065–2070, 2016.

[43] F. Rongqiang, W. Jing, and Y. Zhicheng, ''Computational feature modeling method of multilayer neural network algorithm,'' *Comput. Res. Develop.*, vol. 56, no. 6, pp. 1170–1181,2019.

[44] Y. Xia, Y. Shang, and R. Zhang, ''Structure of PSD-95/MAP1A complex reveals unique target recognition mode of MAGUK GK domain,'' *Biochem. J.*, vol. 474, no. 16, pp. 2817–2828,2017.

[45] E. Holmqvist, L. Li, T. Bischler, L. Barquist, and J. Vogel, ''Global maps of ProQ binding *in vivo* reveal target recognition via RNA structure and stability control at mRNA 3' ends,'' *Mol. Cell*, vol. 70, no. 5, pp. 971–982.e6, Jun. 2018.

[46] H. Göksu, ''Ground moving target recognition using log energy entropy of wavelet packets,'' *Electron. Lett.*, vol. 54, no. 4, pp. 233–235, Feb. 2018.

[47] D. Scarafoni, A. Bockman, and M. Chan, ''Automatic target recognition and geo-location for side scan sonar imagery,'' *J. Acoust. Soc. Amer.*, vol. 141, no. 5, p. 3925, May 2017.

[48] J. M. Carmona and J. Climent, ''Human action recognition by means of subtensor projections and dense trajectories,'' *Pattern Recognit.*, vol. 81, pp. 443–455, Sep. 2018.

[49] M. Majd and R. Safabakhsh, ''A motion-aware ConvLSTM network for action recognition,'' *Int. J. Speech Technol.*, vol. 49, no. 7, pp. 2515–2521, Jul. 2019.

[50] J. Zhang, H. P. H. Shum, J. Han, and L. Shao, ''Action recognition from arbitrary views using transferable dictionary learning,'' *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4709–4723, Oct. 2018.

[51] X. Li, Y. Wang, and G. Liu, ''Structured medical pathology data hiding information association mining algorithm based on optimized convolutional neural network,'' *IEEE Access*, vol. 8, pp. 1443–1452, 2020.

[52] C. Cao, Y. Zhang, C. Zhang, and H. Lu, ''Body joint guided 3-D deep convolutional descriptors for action recognition,'' *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1095–1108, Mar. 2018.

[53] H. T. Nguyen, C. T. Nguyen, P. T. Bao, and M. Nakagawa, ''A database of unconstrained vietnamese online handwriting and recognition experiments by recurrent neural networks,'' *Pattern Recognit.*, vol. 78, pp. 291–306, Jun. 2018.

[54] A. Abdelbaky and S. Aly, ''Human action recognition using short-timemotion energy template images and PCANet features,'' *Neural Comput.*, Jan. 2020, doi: 10.1007/s00521-020-04712-1.

[55] S. Yanhu, Z. Zhang, and H. Kaiqi, ''Review, current situation and Prospect of human visual behavior recognition research,'' *Comput. Res. Develop.*, vol. 53, no. 1, pp. 93–112, 2016.

[56] J.-H. Han, S.-J. Lee, and J.-H. Kim, ''Behavior hierarchy-based affordance map for recognition of human intention and its application to human–robot interaction,'' *IEEE Trans. Human-Machine Syst.*, vol. 46, no. 5, pp. 708–722, Oct. 2016.

WEI LIU received the M.S degree from Northeast Normal University. He is currently an Associate Professor with Heilongjiang University. His research interests include artificial intelligence, sport industry, and sport engineering.

SHANGBIN LI received the M.S. degree from Harbin Engineering University. He is currently a Professor with Harbin Engineering University. His research interests include artificial intelligence and sport engineering.