The Journal of Supercomputing (2021) 77:12192–12208
https://doi.org/10.1007/s11227-021-03772-x

Applying TS-DBN model into sports behavior recognition with deep learning approach

Yingqing Guo[1] · **Xin Wang**[1,2]

Accepted: 22 March 2021 / Published online: 6 April 2021

Abstract

The purposes are to automatically collect information about human sports behavior from massive video data and provide an explicit recognition and analysis of body movements. The analysis of multi-scale input data, the improvement of spatiotemporal Deep Belief Network (DBN), and the different pooling strategies are regarded as the focuses to improve the belief networks in deep learning (DL). Moreover, a human sports behavior recognition model is proposed based on particular spatiotemporal features. Also, video frame data are collected from the Royal Institute of Technology (KTH) and University of Central Florida (UCF) datasets for training. The TensorFlow platform is employed to simulate the built algorithm. Finally, the constructed algorithm model is compared with the DBN proposed by Yang et al. the Convolutional Neural Network (CNN) proposed by Ullah et al. and the DBN-Hidden Markov Model (HMM) algorithm proposed by Xu et al. to analyse its performance. The recognition effects of each algorithm in the two datasets are analyzed. Results demonstrate that CNN developed by Ullah et al. has the worst accuracy on the KTH and UCF datasets, followed by DBN developed by Yang et al. and DBN-HMM developed by Xu et al. The constructed algorithm model can provide the highest accuracy, reaching about 90%, and the recognition accuracy of human sports behaviors of each algorithm on the KTH dataset is lower than that on the UCF dataset. On the KTH dataset, the recognition accuracy for boxing is the highest and running the lowest. Analyzing the model's accuracy in the four scenes (S1, S2, S3, and S4) on the KTH dataset suggests that the recognition accuracy for the indoor scene (S4) is higher than that of the outdoor scenes (S1, S2, and S3). On the UCF dataset, the recognition accuracy for lifting is the highest, reaching 99%, and that for walking is the lowest, reaching 51%. Therefore, the proposed human sports recognition model can provide a higher accuracy than other classic DL algorithms, providing an experimental basis for subsequent human sports recognition research.

Keywords Deep learning · Sports behavior recognition · Deep belief network · UCF dataset · Video

Extended author information available on the last page of the article

1 Introduction

As science and technology develop rapidly and network data grow explosively, obtaining human behavior information from massive video data becomes an urgent issue in many fields. Due to the low efficiency and the constantly-decreasing human attention, long-time manual monitoring of video surveillance often leads to a high loss alarm rate [1]. If intelligent video surveillance is adopted, the video can be automatically modeled and analyzed. Human behaviors can be recognized in real-time for more accurate and in-time security warning, which has broad application prospects in public places, such as transport locations, airport, and stations [2, 3]. Therefore, human behavior recognition has theoretical significances and practical values, becoming the research focus in many fields.

Body actions refer to the simplest limb movements to the entire body's complex joint actions, such as the leg movements when playing football, and the hand, leg, head, and whole-body movements when jumping up and hitting a ball [4]. Body action recognition is often researched from theoretical significance and practical application. In theoretical research, action recognition includes information obtaining and processing. In earlier times, body action information was obtained via some wearable devices. Although the acquired action data were rich, they had significant defects in efficiency, cost, and environment [5]. With video capture equipment's continuous upgrading and updating, body action data can be collected visually. Action recognition based on vision has become a current research hotspot. For example, Kinect, the Time of Flight (TOF) camera developed by Microsoft, can acquire the human body's depth images and joint information, providing significant body action recognition assistance.

To process the body action data, machine-learning algorithms are used to construct models and distinguish new data, such as Support Vector Machines (SVMs), Hidden Markov Model (DBN-HMM), and deep learning (DL) [6]. As for practical application, body language plays a significant role in people's communication. A better understanding of body actions will increase the communication efficiency. Human–computer interaction in this field refers to a machine's understanding of human behaviors through body actions. For example, somatosensory game machines provide a better experience for the players by capturing the players' actions in 3D space.

DL has been widely used in image recognition, classification, evaluation, and predictive analysis in computer vision. It can directly extract information from the original data and form a significant feature expression [7]. First, the original data are preprocessed, and the data features are extracted by hierarchical forward propagation and backpropagation (BP). Each layer's expression is abstracted so that the final expression can better describe the input data [8]. As a DL algorithm, DBN has advantages in action recognition and good modeling capabilities. It can process various input features and establish the connections between adjacent times to extract the actions' context information without assuming the action features' distribution. Therefore, it can be applied in action recognition [9].

In summary, DBN is improved, and a human sports behavior recognition model based on particular spatio-temporal features is proposed to obtain, recognize, and analyse human sports behavior information from massive video data. The constructed algorithm is simulated on Royal Institute of Technology (KTH) and University of Central Florida (UCF) datasets, providing an experimental basis for subsequent sports development and body detection in China.

2 Related works

2.1 Research on body action recognition

Human sports behavior recognition refers to recognizing human behaviors from video sequences. Valuable features can describe various behavior categories, which must be easy to calculate and can respond to the similarity between two similar sports. Many scholars have researched human behavior recognition in kinematics. Patwardhan et al. (2017) proposed a multi-modal emotion recognition method by combining 3D geometric features, kinematic features (joint speed and displacement), and features extracted from daily behavior patterns (such as head point frequency). The 3D geometric and kinematic features were developed by the original feature data in the visual channel, significantly improving human emotions' recognition accuracy [10]. Chiovetto et al. (2018) determined dynamic facial expressions' adequate dimensions by learning the collected facial expressions. The Bayesian model simulated different numbers of primitive models, finding that a few independent control units might control facial expressions, allowing facial expressions' low-dimensional parameterization [11]. Yang et al. (2019) proposed a multi-sensor integrated system and a two-level activity recognition classifier to assist rehabilitation exercises, finding out that the system's accuracy was much improved and could predict falling time and direction, as well as abnormal gait types [12]. Hu et al. (2020) proposed a network structure combining the batch normalization algorithm with the GoogLeNet network model to solve the problems of complicated action feature extraction and low recognition accuracy and improve the algorithm's performance in body action recognition. The results showed that the improved DL algorithm significantly improved recognition accuracy and body recognition advantages [13].

3 Research on DL's application trend

With the rapid development of science and technology, the big data era has arrived. DL has been applied in various fields. Ohsugi et al. (2017) applied DL in material medicine, using ultra-wide-field fundus images to detect the Rhegmatogenous Retinal Detachment (RRD). They found that ophthalmology clinics' medical services in remote areas were significantly improved [14]. Sremac et al. (2018) established an online shopping management system applicable to various supply chain goods with high accuracy [15]. Wu et al. (2019) improved disease treatments by DL algorithms and understanding the patient's physical conditions [16]. Ghosh et al. (2019)

proposed a DL method of molecular excitation spectrum prediction. Three different neural network structures, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Deep Tensor Neural Network (DTNN), were trained and evaluated to analyse organic molecules' electronic density of states. They discovered that the proposed method could predict the tiny organic molecules' structure in real-time and determine the potential application molecules [17]. Shao et al. (2020) developed a CNN-based predictor of viral protein subcellular positioning for infectious diseases caused by a coronavirus, COVID-19, and H1N1 currently spreading around the world, called Ploc-Deep-mVirus. They found that this predictor was particularly suitable for processing multi-site systems, and its predictive performance was significantly better than that of advanced predictive indicators at present [18].

Although DL has been widely applied in various fields, its application in body action recognition is scarce. The deep model's structure is complex, and overfitting may occur in model parameters learning. Therefore, DL is improved for human sports behavior recognition to promote model learning's robustness, which is significant for developing human–computer interaction and sports analysis.

4 Method

4.1 Body action recognition and analysis

Due to perspective changes, different actions may generate similar projections in human behavior recognition and analysis. Various environmental factors, such as illumination changes and mutual covering, make human behavior recognition uneasy. Therefore, as the primary contents of human behavior recognition, the selection of features and effective descriptors from the video image sequences to describe the body movement state can reduce the spatio-temperal dimensionalities and the calculation complexity [19]. Sports behaviors are characterized by selecting appropriate features, and a classifier is trained to classify human actions by machine learning (ML) methods, obtaining the final recognition results. Figure 1 shows the process of human sports behavior recognition [20]. ML's primary

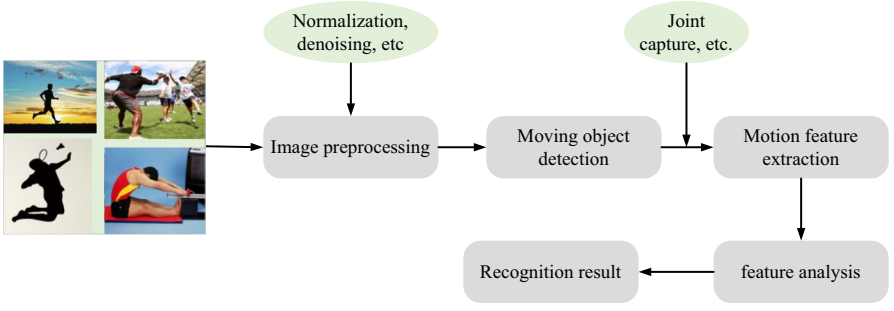

Fig. 1 Human sports behavior recognition

functions are to reduce feature extraction complexity and improve image feature discrimination and behavior recognition's robustness in this process.

Human action recognition becomes a matter of classification when images can represent image frames or sequences. Logistic regression classification [21], softmax regression classification [22], naive Bayes [23], and SVM [24] are common human action recognition classification methods. The softmax classification is used herein. Logistic regression classification is suitable for two-category classification. However, multi-category classification is more common. There are usually two choices for this k binary classifiers and multi-classifier softmax regression extended to logistic regression. If there is a multi-category classification issue at present expressed as $y^{(i)} \in \{1, 2, ..., k\}$, with a total of k categories. For a given test x, Eq. (1) shows category probability assumed in softmax regression classification.

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1|x^{(i)};\theta\right) \\ p\left(y^{(i)} = 2|x^{(i)};\theta\right) \\ \vdots \\ p\left(y^{(i)} = k|x^{(i)};\theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{1}$$

θ represents the model parameters, a matrix of k lines. Each line can be regarded as a category's classifier parameter, as recorded in Eq. (2).

$$\theta = \begin{bmatrix} \theta_1^t \\ \theta_2^t \\ \vdots \\ \theta_k^t \end{bmatrix} \tag{2}$$

$\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}$ normalizes the probability distribution so that the sum of all probabilities is one. Equation (3) shows the system's cost function equation.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(j)} = j\right\}\right] \log \frac{\theta_j^T x^{(i)}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \tag{3}$$

For the indicative function $1\{\cdot\}$, the value rules are *1{expression for true value} = 1 and 1{expression for false value} = 0.* Then, Softmax regression accumulates k categories' probabilities. Equation (4) shows the probability that x is classified into j categories.

$$\log p\left(y^{(i)} = j|x^{(i)};\theta\right) = \frac{\theta_j^T x^{(i)}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \tag{4}$$

Equation (3) is the cost function generalization of logistic regression. Equation (5) shows the regression cost function.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(j)} = j\right\} \log p\left(y^{(i)} = j|x^{(i)};\theta\right)\right] \tag{5}$$

Similarly, an iterative optimization algorithm can minimize the cost function in this equation, such as a gradient descent method. Therefore, Eq. (6) shows the calculation of the loss function's partial derivative.

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[x^{(i)}\left(1\left\{y^{(j)} = j\right\} - p\left(y^{(i)} = j|x^{(i)};\theta\right)\right)\right] \tag{6}$$

In (6), $\nabla_{\theta_j} J(\theta)$ represents a vector, and its l-th $\frac{\partial J(\theta)}{\partial \theta_{jl}}$ represents the l-th partial derivative in the cost function's j-th category. The above equation is substituted into the gradient descent algorithm and iteratively updated to minimize the cost function. With the same number subtracted from each obtained optimal parameter, the loss function's value obtained does not change, indicating that the parameter is not the only solution. Equation (7) shows the proof process.

$$p\left(y^{(i)} = j|x^{(i)};\theta\right) = \frac{e^{(\theta_t - \psi)_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{(\theta_t - \psi)_j^T x^{(i)}}} = \frac{e^{\theta_j^T x^{(i)}} e^{-\psi_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}} e^{-\psi_j^T x^{(i)}}} = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}} \tag{7}$$

Weight attenuation is added to the cost function to punish excessive parameter values and ensure that the cost function is the strictest convex function. Converging to the optimal global solution, Eq. (8) shows the cost function.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(j)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \tag{8}$$

In (8), $\lambda > 0$. Equation (9) shows the partial derivative function.

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{j=1}^{m}\left[x^{(i)}\left(1\left\{y^{(j)} = j\right\} - p\left(y^{(i)} = j|x^{(i)};\theta\right)\right)\right] + \lambda\theta_j \tag{9}$$

Finally, a usable softmax regression classification model is obtained by minimizing the cost function.

4.2 DBN feature learning and analysis

In recent years, unsupervised feature learning receives widespread attention in computer vision because automatic learning to obtain robustness' visual feature expression in the massive unlabeled data (including images and videos) becomes a crucial task for the next generation of intelligent vision applications [25]. In the ML application, computer vision and neuroscience researchers have reached the consensus shown in Fig. 2 in the feature extraction and unsupervised feature learning.

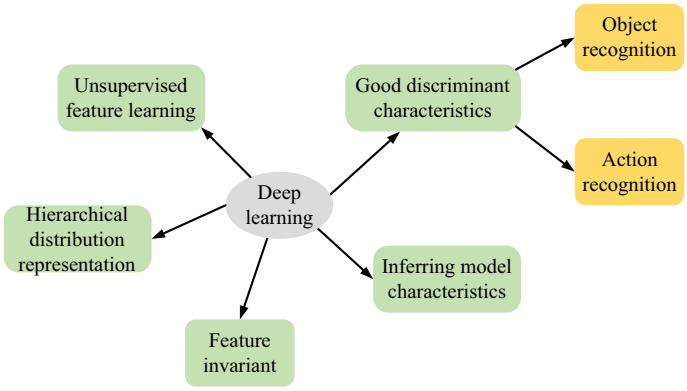

Fig. 2 DL's consensus about feature extraction and learning in DL

DL application in action recognition adopts a process of video preprocessing, feature detection, feature description, and classification. First, the video data are preprocessed to reduce factors unrelated to the recognition. Second, the spatio-temporal feature detector selects a local interest region, significantly reducing the data amount that needs to be considered. After obtaining the salient points, the feature descriptors adopt the bag-of-words model and discard all position information. Finally, after the descriptor's calculation, the features are sent to the classifier, K-Nearest Neighbor (KNN) algorithm classifier [26], or SVM classifier. Feature descriptors describe the movements and appearance characteristics in salient points' the adjacent areas.

As a fundamental DBN module, Restricted Boltzmann Machine (RBM) is a bidirectional hidden-variable model, including a set of visible nodes V and a set of hidden nodes h. The two sets are not connected internally, but the nodes are entirely connected with the connection weight represented by W. A real-valued displacement is added to each node, with displacements of V and h represented by b and d, respectively. The parameter set is represented by φ, including $W \in R^{n_v \times n_H}, b \in R^{n_H}$, and $d \in R^{n_V}$, where n_v represents the number of visible nodes, and n_H denotes the number of hidden nodes. The model's bias toward V, and h is often represented by defining the energy equation $E(v,h,\varphi)$, where the lower the energy value, the more biased the model toward the node pair of V and h. When φ is known, the joint distribution of node pairs V and h can be represented by the all possible node pairs' energy equation after exponential normalization. Equations (10) to (13) show the integral formula on h.

$$p(v, h|\varphi) = \frac{1}{z(\varphi)} \exp(-E(v, h, \varphi)) \tag{10}$$

$$z(\varphi) = \int_{V' \in V, h' \in H} \exp(-E(v, h, \varphi)) \tag{11}$$

$$p(v|\varphi) = \frac{\exp(-F(v,\varphi))}{\int_{V' \in V} \exp(-F(v,\varphi))} \tag{12}$$

$$F(v,\varphi) = -\int_{h \in H} \exp(-E(v,h,\varphi)) \tag{13}$$

$z(\varphi)$ represents the normalization constant, $F(v, \varphi)$ denotes the free energy, and v and H signify the dimensions of the variables v and h. H is only limited to a Boolean quantity, $H=\{0,1\}^n{}_H$, and V can be a Boolean quantity or a continuous quantity. Further variants of the RBM model generates the convolution restricted Boltzmann machine (CRBM) [27], including three sets of nodes, the visible layer node v, the hidden layer node h, and the pooling layer node p. Maximum probability pooling is usually adopted in the pooling layer, activated only when at least one of its corresponding hidden layer nodes is activated. Figure 3 shows the CRBM calculation process.

The CRBM model can recognize the repeated local features from the images, and mandatory displacement invariance is realized in the model. However, the CRBM model assumes that the images are independently distributed, not applicable to video modeling's time structure due to the video frames' relevance [28]. Therefore, this model is further improved by modeling separately in time and space, in turn, to make it more invariant in spatio-temporal transformation. A hierarchical method is adopted herein, that is, the distributed probability model, which learns the spatio-temporal invariant features from the videos using unsupervised learning. Specifically, as the introductory module, the CRBM learns the original data's hierarchical structure. It has an increasingly complex structure from the bottom to the top, and the invariance gradually increases, called the spatio-temporal Deep Belief Network (ST-DBN). In this model, repeated operations are performed in the time and the

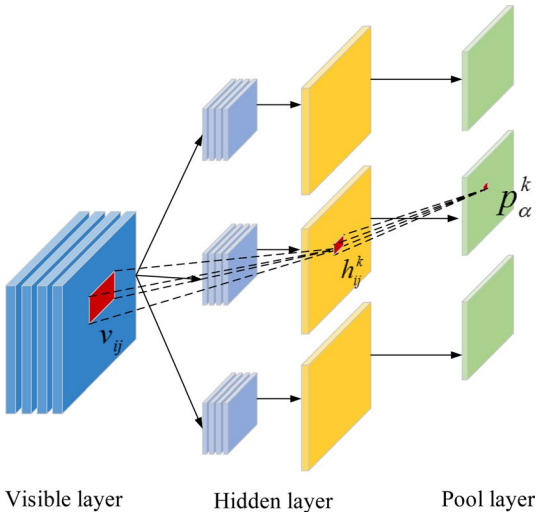

Fig. 3 The calculation process of CRBM

space dimensions successively so that the upper layers of the network can maintain characteristics invariance in broader spatio-temporal dimensions.

4.3 Construction of human sports behavior recognition model based on sparse spatio-temporal features

To recognize and analyse human behaviors from videos, the improved ST-DBN can learn sparse spatio-temporal features to recognize human sports behaviors. Figure 4 shows the human sports behavior recognition model based on sparse spatio-temporal features.

In the multi-scale spatial expression, spatio-temporal Gabor [29] is used in the input videos to convolve with the original input video and construct the scale space. The model training complexity and the information loss among different scale expressions are considered; three scales of minimum losses are selected as the input videos' multi-scale expression to input the deep model and learn multi-scale features.

4.3.1 Learning of sparse spatio-temporal features

When learning the sparse spatio-temporal features of human sports behaviors, different scale expressions are used as TS-DBN's different channel values to jointly learn multi-scale features and the information interaction between different scales. The traditional ST-DBNCRBM learns features in spatio-temporal dimensions separately with spatial CRBM as the first layer and temporal CRBM as the second layer, which are stacked in sequence for automatic spatio-temporal feature learning. However, behavior evolution in the time dimension is more significant than that in the space dimension. For example, for running and trotting behavior categories, changes in the space dimension are inconspicuous but significant in the time dimension. Therefore, the features in the time dimension should be learned first in behavior recognition. The S-T DBN first performs CRBM in the time dimension and then in

temporal features.

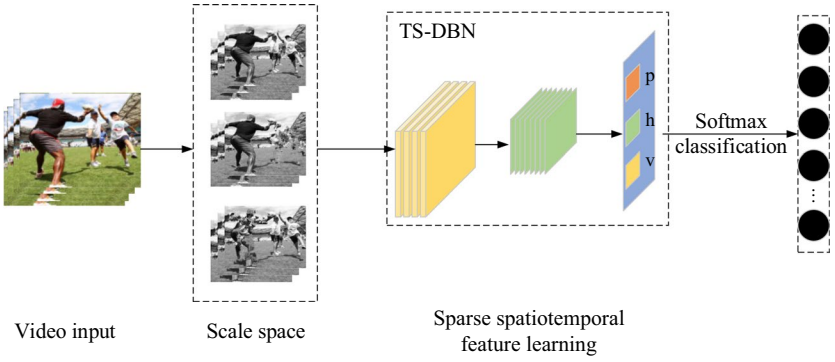

Fig. 4 Human sports behavior recognition model based on sparse spatio-temporal features

the space dimension to learning spatio-temporal features, called Time–Space Deep Belief Network (TS-DBN).

Specifically, the multi-scale TS-DBN uses different scale expressions of ST-DBN's input videos as the different channels' values to jointly learn spatio-temporal features in different scales. The CRBM's input in the time dimension is the pixel's vector at the position *(i, j)* in the image in the time dimension, that is, a time sequence with a length of $(ch \times nV \times 1)$, with *ch* as the video channels' number and different information scales and *nV* as the video length. During learning, CRBM model outputs the $(|w| \times nT \times 1)$'s sequence with $|w|$ as the filters' number and *nT* the output video's length. Finally, the time dimension output is rearranged in the space dimension distribution.

4.3.2 TS-DBN algorithm training

The improved TS-DBN model is trained through greedy hierarchical pre-training. From its lowest layer, each model's input layers are trained randomly. Then, after the hidden layer expression, it is rearranged and input to the next layer. This process is repeated continuously throughout the training until all layers' training is completed. After the entire network is trained, the layer's hidden node expressions can be extracted from any given layer in the video.

Conversely, with given hidden node expressions, the video samples can be calculated. The maximum merging unit's feedback probabilities are calculated in each layer to extract its features. The continuous probabilities of hidden nodes and maximum merged nodes approximate their posterior probabilities.

For sampling, the hidden node and maximum merging layer are initialized to the average value. After the Gibbs sampling from the top layer down [30], the sampled values are passed backward from top to bottom through BP. Equations (14) and (15) show that the hidden nodes' conditional probabilities are obtained by the probability integral's uniform distribution in each layer.

$$P(HP_a^g = -p_a^g | h\prime) = 1 - P\left(p_a^g | h\prime\right) \tag{14}$$

$$P(HP_a^g = h_{r,s}^g | h\prime) = \frac{1}{\left|B_a^g\right|} P\left(p_a^g | h\prime\right) \tag{15}$$

In Eqs. (14) and (15), *h'* represents the hidden nodes in the current layer and the previous layer, and $P\left(p_a^g | h'\right)$ presents p_a^g's top-down belief value. Due to the resolution reduction caused by maximum merging, the top-down information cannot produce precisely consistent detailed information with the bottom-up input. The entire network is sampled by Gibbs sampling until convergence.

4.3.3 Human sports behavior recognition dataset

A public dataset is used to analyse the human sports behavior classification model. Common public datasets include Weizmann [31], KTH [32], and UCF [33]. KTH and UCF are adopted for simulation. About 599 videos in the KTH dataset include

Fig. 5 Sample frames of KTH action database (in the first row) and UCF sports database (in the second and third rows)

Table 1 Specific experimental environment configurations

		Versions
Software	Operating system	Linux 64bit
	Python version	Python 3.6.1
	TensorFlow version	1.0.1
Hardware	CPU	Intel Core i7-7700@4. 0 GHz 8 cores
	Storage	Kingston ddr4 2400 MHz 16G
	GPU	Nvidia GeForce 1060 8G

six behaviors: jogging, walking, boxing, running, hand waving, and hand clapping. Each behavior is performed by 25 people in four different environmental scales, including S1 (outdoor environment with constant scale), S2 (outdoor environment with varying scale), S3 (outdoor environment with different clothes), and S4 (indoor environment with varying lighting). There are 150 sports videos in the UCF dataset, including diving, kicking, walking, golf, lifting, running, skateboard, swinging, Swing-Side Angle, and horse riding. Most of the performers' appearances in this dataset are quite different. The background is also noisy, and the lighting conditions change significantly due to the camera movements. Figure 5 shows the example frames of the KTH action database and UCF sports database.

4.4 Entity resolution

Absorbing other platforms' advantages, the TensorFlow platform [34] simulates the human sports behavior recognition model with sparse spatio-temporal features. It has developed into a mature and complete DL framework with installation versions, such as Windows, Linux, and Mac OS X. Table 1 summarizes the experimental

environment configurations. After the TensorFlow platform is installed, a Python terminal can be opened for testing. Then, the image data in the KTH and UCF datasets are collected, and the performance of the model is analyzed. In the parameter setting, the bias value is initialized to 0, and the weight W is initialized with a random value from the normal distribution $N(0, 0.01)$. To accelerate the learning efficiency, momentum is introduced and initialized to 0.5. Samples of each batch are selected randomly, and the size of mini-batches is doubled. The algorithm model is compared with the research of other scholars, including the DBN developed by Yang et al. [35], CNN developed by Ullah et al. [36], and DBN-HMM developed by Xu et al. [37].

5 Results and discussion

5.1 The recognition effect analysis of various algorithms in the two datasets

Figure 6 indicates the comparison and analysis of the human sports behavior recognition model proposed with the DBN developed by Yang et al. CNN developed by Ullah et al. and DBN-HMM developed by Xu et al. CNN's accuracy is the lowest on the KTH and UCF datasets, followed by DBN and DBN-HMM. The proposed TS-DBN algorithm can provide the highest accuracy. These algorithms' accuracies are higher on KTH dataset than that on the UCF dataset. Therefore, the above results infer that the proposed algorithm model's accuracy is higher than that of the traditional CNN and DBN algorithms. The reason is that the human kinematics characteristics are well extracted, and the TS-DBN algorithm model eliminates boundary noises. The cameras and the complex backgrounds are removed, and the feature information is abstracted many times to obtain high-level features, further enhancing the describing ability for actions and detailed spatial information. Thus, its final accuracy rate is significantly higher than that of other methods.

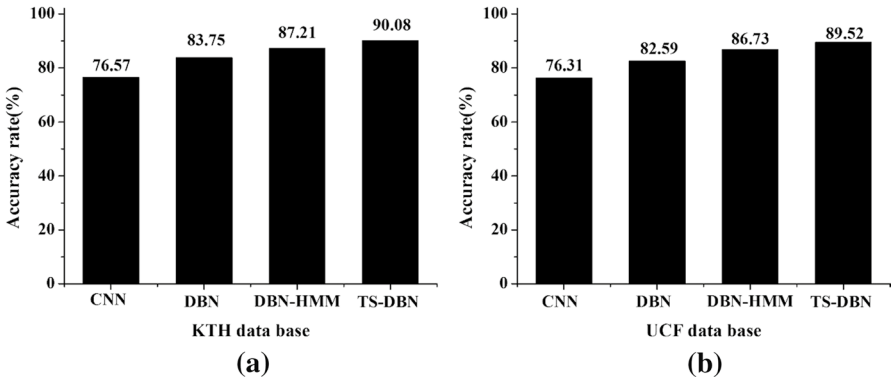

Fig. 6 Each algorithm's recognition accuracy on the two datasets (**a**: on the KTH dataset, **b**: on the UCF dataset)

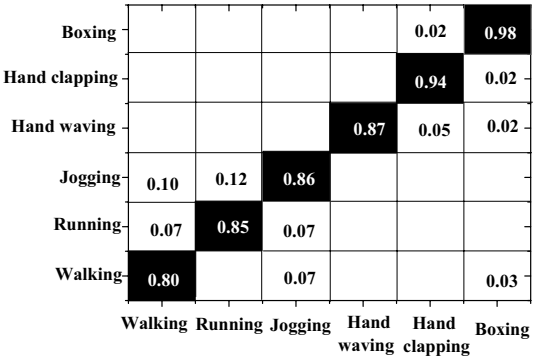

Fig. 7 Confusion matrix of various behavior recognition on the KTH dataset

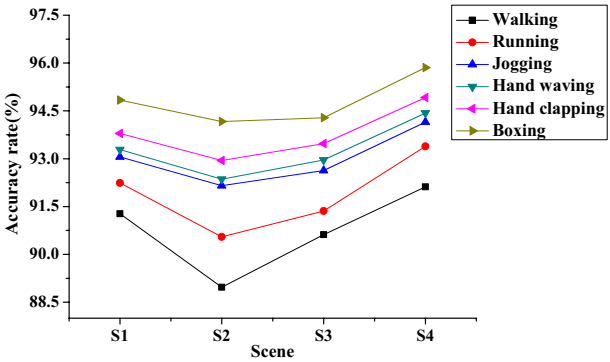

Fig. 8 Recognition accuracy for various behaviors on the KTH dataset in four different scenes

5.2 Accuracy results and analysis on the KTH dataset

Figure 6 indicates the comparison and analysis of the proposed human sports behavior recognition model with the DBN developed by Yang et al. CNN developed by Ullah et al., and DBN-HMM developed by Xu et al. CNN's accuracy is the lowest on the KTH and UCF datasets, followed by DBN and DBN-HMM. The proposed TS-DBN algorithm can provide the highest accuracy. These algorithms' accuracies are higher on the KTH dataset than that on the UCF dataset. Therefore, the above results infer that the proposed algorithm model's accuracy is higher than that of the traditional CNN and DBN algorithms. The reason is that the human kinematics characteristics are well extracted, and the TS-DBN algorithm model eliminates boundary noises. The cameras and the complex backgrounds are removed, and the feature information is abstracted many times to obtain high-level features, further enhancing the describing ability for actions and detailed spatial information. Thus, its final accuracy rate is significantly higher than that of other methods (Fig. 7).

Figure 8 shows the action recognition accuracy analysis in the four scenes (S1, S2, S3, and S4) on this dataset. The indoor scene's (S4) recognition result is

better than that of the outdoor scenes (S1, S2, and S3). The reason is that people's actions in outdoors are easily affected by illumination. The recognition accuracy rate is the lowest in the S2 scene. Although the S2 scene is affected by illumination, there are angle and scale changes caused by the lens expansion. However, accuracy differences in these scenes are slight, showing that the proposed TS-DBN network performs excellently in different scenes.

5.3 Accuracy results and analysis on the UCF dataset

Figure 9 shows the classification confusion matrix of ten behaviors in UCF by analyzing the accuracy in the UCF sports database, indicating that the proposed method has reasonable accuracy rates. Besides, the accuracy rate for lifting is the highest and walking the lowest, reaching 99% and 51%, respectively. There are misclassifications for similar behaviors, such as kicking and running.

6 Discussion

With the development of artificial intelligence technology, human action recognition has become a research hotspot in computer vision and pattern recognition, which has attracted widespread attention from scholars in various fields. Here, a human sports behavior recognition model is proposed by improving DBN based on particular spatio-temporal characteristics. This model is then compared with DBN developed by Yang et al. CNN developed by Ullah et al. and DBN-HMM developed by Xu et al. on the KTH and UCF datasets. The proposed TS-DBN can provide the best effects of human sports behavior recognition, followed by DBN-HMM developed by Xu et al. CNN developed by Ullah et al. has the worst recognition effect. A possible reason is that the constructed TS-DBN algorithm model captures the human kinematics characteristics well, and simultaneously eliminates boundary noise, removes the camera and complex background, and abstracts the feature information many times to obtain high-level features. Hence, its ability to describe movement information

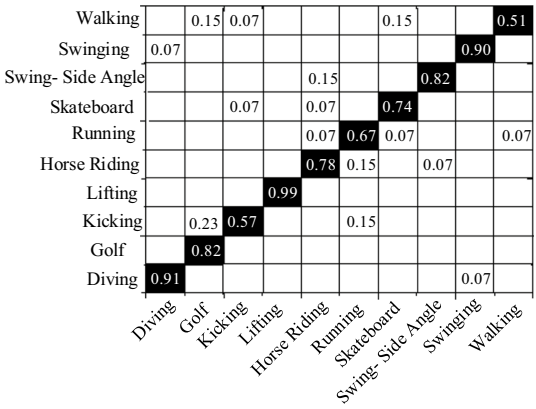

Fig. 9 Confusion matrix of various behaviors in the UCF dataset

and detailed spatial information is further enhanced so that the final accuracy rate is significantly higher than other methods.

Furthermore, the constructed algorithm model's accuracy is analyzed from the two datasets of KTH and UCF. The accuracies of recognizing actions in the KTH dataset are analyzed. The results reveal that the action with the highest recognition accuracy is punching, and that with the lowest recognition accuracy is running (80%). Analyzing the recognition accuracy for the four scenes (S1, S2, S3, and S4) on the dataset finds that the recognition accuracy of the indoor scene (S4) is significantly better than that of the three outdoor scenes (S1, S2, and S3). On the UCF dataset, lifting has the highest recognition accuracy rate, reaching 99%, and walking has the lowest recognition accuracy rate, only 51%. This result shows that the proposed sports recognition model based on the TS-DBN algorithm is helpful on different datasets. The average accuracy rate is also better than that of the algorithm model proposed by other scholars. Finally, the robustness and effectiveness of the proposed TS-DNB algorithm in human sports behavior recognition are confirmed.

7 Conclusion

DL application focuses on analyzing multi-scale input data, improving spatio-temporal DBN, and exploring different pooling strategies. Here, a TS-DBN algorithm is proposed for human sports behavior recognition based on DL. The simulation shows that on the KTH and UCF datasets, the recognition accuracy of the constructed model is higher, reaching about 90%, which is better than the recognition accuracy of models proposed by other scholars. In the meantime, the model is effective on different datasets, which can provide an experimental basis for recognizing human sports in the future.

However, there are some shortcomings as well. First, only the brightness information is used for the input video, but the color information is not considered. Usually, color information contains many features. For some behaviors or other applications, learning features from color space input is more conducive to improving the recognition rate. Besides the color features, some other features can also be input and integrated with the DL model to improve the behavior recognition rate. The proposed model has reduced the calculation amount; however, some pretreatments are required compared to the original input's direct processing. Therefore, whether the feature vector provided by preprocessing is good enough is an issue that can be further improved.

References

1. Fuentes A, Yoon S, Park J, Park DS (2020) Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. Comput Electron Agric 177:105627
2. Pan MK, Skjervøy V, Chan WP, Inaba M, Croft EA (2017) Automated detection of handovers using kinematic features. Int J Robot Res 36(5–7):721–738
3. Al-Janabi S, Salman AH (2021) Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications. Big Data Min Anal 4(2):124–138

4. Kim H, Lee S, Kim Y, Lee S, Lee D, Ju J, Myung H (2016) Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system. Expert Syst Appl 45:131–141
5. Dingenen B, Staes FF, Santermans L, Steurs L, Eerdekens M, Geentjens J, Deschamps K (2018) Are two-dimensional measured frontal plane angles related to three-dimensional measured kinematic profiles during running? Phys Ther Sport 29:84–92
6. Aurangzeb K, Haider I, Khan MA, Saba T, Javed K, Iqbal T, Sarfraz MS (2019) Human behavior analysis based on multi-types features fusion and Von Nauman entropy based features reduction. J Med Imaging Health Inf 9(4):662–669
7. Edey R, Yon D, Cook J, Dumontheil I, Press C (2017) Our own action kinematics predict the perceived affective states of others. J Exp Psychol Hum Percept Perform 43(7):1263
8. Balasundaram A, Pradeep KV, Sandhya S (2021) An Extensive study on disease prediction in mango trees using computer vision. Ann Rom Soc Cell Biol 25(1):1895–1905
9. Wan S, Qi L, Xu X, Tong C, Gu Z (2020) Deep learning models for real-time human activity recognition with smartphones. Mob Netw Appl 25(2):743–755
10. Patwardhan A (2017) Three-dimensional, kinematic, human behavioral pattern-based features for multimodal emotion recognition. Multimodal Technol Interact 1(3):19
11. Chiovetto E, Curio C, Endres D, Giese M (2018) Perceptual integration of kinematic components in the recognition of emotional facial expressions. J Vis 18(4):13–13
12. Yang T, Gao X, Gao R, Dai F, Peng J (2019) A novel activity recognition system for alternative control strategies of a lower limb rehabilitation robot. Appl Sci 9(19):3986
13. Hu Z, Park SY, Lee EJ (2020) Human motion recognition based on spatio-temporal convolutional neural network. J Korea Multimed Soc 23(8):977–985
14. Al-Janabi S, Alkaim AF, Adel Z (2020) An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft Comput 24(14):10943–10962
15. Sremac S, Tanackov I, Kopić M, Radović D (2018) ANFIS model for determining the economic order quantity. Decis Mak Appl Manag Eng 1(2):81–92
16. Wu Y, Luo Y, Chaudhari G, Rivenson Y, Calis A, De Haan K, Ozcan A (2019) Bright-field holography: cross-modality deep learning enables snapshot 3D imaging with bright-field contrast using a single hologram. Light Sci Appl 8(1):1–7
17. Ghosh K, Stuke A, Todorović M, Jørgensen PB, Schmidt MN, Vehtari A, Rinke P (2019) Deep learning spectroscopy: neural networks for molecular excitation spectra. Adv Sci 6(9):1801367
18. Shao Y, Chou KC (2020) pLoc_Deep-mVirus: a CNN model for predicting subcellular localization of virus proteins by deep learning. Nat Sci 12(6):388–399
19. Jalal A, Mahmood M (2019) Students' behavior mining in e-learning environment using cognitive processes with information technologies. Educ Inf Technol 24(5):2797–2821
20. Corrigan BW, Gulli RA, Doucet G, Martinez-Trujillo JC (2017) Characterizing eye movement behaviors and kinematics of non-human primates during virtual navigation tasks. J Vis 17(12):15–15
21. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J Neurosci 38(33):7255–7269
22. Hassan MM, Uddin MZ, Mohamed A, Almogren A (2018) A robust human activity recognition system using smartphone sensors and deep learning. Futur Gener Comput Syst 81:307–313
23. Prati A, Shan C, Wang KIK (2019) Sensors, vision and networks: from video surveillance to activity recognition and health monitoring. J Ambient Intell Smart Environ 11(1):5–22
24. Al-Janabi S, Mohammad M, Al-Sultan A (2020) A new method for prediction of air pollution based on intelligent computation. Soft Comput 24(1):661–680
25. Vieira ST, Rosa RL, Rodríguez DZ (2020) A speech quality classifier based on tree-CNN algorithm that considers network degradations. J Commun Softw Syst 16(2):180–187
26. Al-Janabi S, Alkaim AF (2020) A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft Comput 24(1):555–569
27. Al-Janabi S (2020) Smart system to create an optimal higher education environment using IDA and IOTs. Int J Comput Appl 42(3):244–259
28. Chen Y, Xu W, Zuo J, Yang K (2019) The fire recognition algorithm using dynamic feature fusion and IV-SVM classifier. Clust Comput 22(3):7665–7675

29. Alkaim AF, & Al-Janabi S (2019). Multi objectives optimization to gas flaring reduction from oil production. In: International Conference on Big Data and Networks Technologies, Springer, Cham, p 117–139
30. Ali SH (2012). Miner for OACCR: case of medical data analysis in knowledge discovery. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT). IEEE, p 962–975
31. Huan RH, Xie CJ, Guo F, Chi KK, Mao KJ, Li YL, Pan Y (2019) Human action recognition based on HOIRM feature fusion and AP clustering BOW. PLoS ONE 14(7):e0219910
32. Jaouedi N, Boujnah N, Bouhlel MS (2020) A new hybrid deep learning model for human action recognition. J King Saud Univ-Comput Inf Sci 32(4):447–453
33. Perera AG, Law YW, Chahl J (2019) Drone-action: an outdoor recorded drone video dataset for action recognition. Drones 3(4):82
34. Mahdi MA, & Al_Janabi S (2019). A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. In: International Conference on Big Data and Networks Technologies, Springer, Cham, p 320–339
35. Yang H, Yuan C, Li B, Du Y, Xing J, Hu W, Maybank SJ (2019) Asymmetric 3d convolutional neural networks for action recognition. Pattern Recogn 85:1–12
36. Ullah A, Muhammad K, Haq IU, Baik SW (2019) Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Futur Gener Comput Syst 96:386–397
37. Xu J, & Luo Q (2021). Human action recognition based on mixed gaussian hidden markov model. In: MATEC Web of Conferences, Vol 336, EDP Sciences, p 06004

Authors and Affiliations

Yingqing Guo[1] · Xin Wang[1,2]

✉ Xin Wang
369012381@qq.com

Yingqing Guo
1274611020@qq.com

[1] Institute of Physical Education, Shandong University, Jinan, China

[2] College of Physical Education, Liaocheng University, Liaocheng, China